UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

"Detective Work We Shouldn't Have to Do"

Practitioner Challenges in Regulatory-Aligned Data Quality in Machine Learning Systems

Wang, Yichun; Irion, Kristina; Groth, Paul; Harmouch, Hazar

DOI
10.1145/3805689.3812313

Publication date
2026

Document Version
Final published version

Published in
ACM FAccT 2026

License
CC BY

Link to publication

Citation for published version (APA):
Wang, Y., Irion, K., Groth, P., & Harmouch, H. (2026). "Detective Work We Shouldn't Have to Do": Practitioner Challenges in Regulatory-Aligned Data Quality in Machine Learning Systems. In *ACM FAccT 2026: Proceedings of the 9th annual ACM Conference on Fairness, Accountability, and Transparency : June 25-28, 2026, Montréal, Canada* (pp. 3217-3241). Association for Computing Machinery. https://doi.org/10.1145/3805689.3812313

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:17 Nov 2026

"Detective Work We Shouldn't Have to Do": Practitioner Challenges in Regulatory-Aligned Data Quality in Machine Learning Systems

YICHUN WANG, University of Amsterdam, The Netherlands
KRISTINA IRION, University of Amsterdam, The Netherlands
PAUL GROTH, University of Amsterdam, The Netherlands
HAZAR HARMOUCH, University of Amsterdam, The Netherlands

Ensuring data quality in machine learning (ML) systems has become increasingly complex as regulatory requirements expand. In the European Union (EU), frameworks such as the General Data Protection Regulation (GDPR) and the Artificial Intelligence Act (AI Act) articulate data quality requirements that closely parallel technical concerns in ML practice, while also extending to legal obligations related to accountability, risk management, and human rights protection. This paper presents a qualitative interview study with EU-based data practitioners working on ML systems in regulated contexts. Through semi-structured interviews, we investigate how practitioners interpret regulatory-aligned data quality, the challenges they encounter, and the support they identify as necessary. Our findings reveal persistent gaps between legal principles and engineering workflows, fragmentation across data pipelines, limitations of existing tools, unclear responsibility boundaries between technical and legal teams, and a tendency towards reactive, audit-driven quality practices. We also identify practitioners' needs for compliance-aware tooling, clearer governance structures, and promoting a culture of regulatory-aligned data quality.

CCS Concepts: • **Social and professional topics** → **Governmental regulations**; **Quality assurance**; • **Human-centered computing** → **User studies**; • **Computing methodologies** → *Machine learning*.

Additional Key Words and Phrases: Data Quality, Regulatory Requirement, Compliance, AI Act, GDPR, Practitioners, European Union

ACM Reference Format:
Yichun Wang, Kristina Irion, Paul Groth, and Hazar Harmouch. 2026. "Detective Work We Shouldn't Have to Do": Practitioner Challenges in Regulatory-Aligned Data Quality in Machine Learning Systems. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3805689.3812313

1 Introduction

Data quality plays a central role in the reliability, fairness, and safety of ML systems [25, 35, 65]. Poor-quality data can lead to biased outcomes, unstable models, and system failures that affect individuals and organisations alike [52, 60, 70]. In much of the technical research, data quality in ML is framed in terms of performance and robustness. Practitioners focus on avoiding label noise, distribution shift, data drift, and inconsistencies that degrade model accuracy or stability [10, 56, 58]. To this end, they deploy validation checks, monitoring, and pipeline orchestration tooling to keep pipelines stable in production [69, 82].

Beyond these technical considerations, in the EU, data quality is also a regulatory requirement tied to accountability, risk management, and the protection of fundamental rights. EU regulations such as the General

Authors' Contact Information: Yichun Wang, University of Amsterdam, The Netherlands; Kristina Irion, University of Amsterdam, The Netherlands; Paul Groth, University of Amsterdam, The Netherlands; Hazar Harmouch, University of Amsterdam, The Netherlands.

This work is licensed under a Creative Commons Attribution 4.0 International License.
FAccT '26, Montreal, QC, Canada
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3812313

Data Protection Regulation (GDPR) and the newly enacted Artificial Intelligence Act (AI Act) introduce explicit requirements related to data quality [22, 23]. The GDPR establishes principles such as accuracy, data minimisation, purpose limitation, and accountability in relation to the processing of personal data [29, 36]. Once it becomes applicable, the AI Act further requires appropriate and representative training, validation, and testing data, bias mitigation measures, documentation, and ongoing risk management for high-risk AI systems [46, 53, 75]. Together, these regulations require data quality practices that satisfy both technical and legal demands.

In practice, aligning data quality with regulatory requirements is not straightforward. Modern ML systems rely on complex pipelines that span data ingestion, transformation, feature engineering, training, deployment, and monitoring [10, 59, 64, 70]. Responsibilities for data quality are distributed across teams, tools, and organisational boundaries [49, 57, 65]. Legal requirements are expressed as high-level principles, while engineers work with concrete constraints, metrics, and delivery timelines [5, 20, 57]. This complicates the operationalisation of regulatory requirements into concrete data quality practice across multiple pipeline components and organisational contexts. To ground this challenge, we conceptualise "regulatory-aligned data quality" as the extent to which data practices meet engineering robustness expectations while also being structured, documented, and governed to demonstrably satisfy applicable EU legal obligations, including traceability, accountability, bias mitigation, and auditability. It differs from general responsible AI practices in that it is grounded in enforceable legal obligations rather than aspirational ethical principles, and therefore requires demonstrable, evidence-based traceability rather than normative alignment alone. To operationalise this concept empirically, we systematically map commonly used data quality dimensions to the corresponding requirements in the GDPR and the AI Act (see Appendix A, Table 3) [80]. It serves both as the analytic lens for our coding and as the organising structure for the interview protocol (see Appendix B and Appendix D, Table 4). However, how practitioners operationalise this alignment in real-world ML pipelines remains underexplored. As a result, practitioners often face uncertainty about how regulatory requirements should be interpreted and operationalised in everyday data quality assurance practices [44, 62, 72, 77]. Understanding how they navigate this space is essential, because their lived experience reveals where structural, technical, and organisational support is most needed, and provides an empirical basis for designing better tools and governance models that neither technical nor legal scholarship can produce alone.

This paper addresses this gap by examining regulatory-aligned data quality from the perspective of practitioners. Practitioners play a central role in operationalising regulatory requirements through their day-to-day data work, yet their perspectives on regulatory-aligned data quality remain underexplored. We ask the following research question:

What real-world challenges do practitioners face in ensuring regulatory-aligned data quality in ML systems, and what forms of support do they need to address these challenges in practice?

To effectively address this central question, the study will delve into the following specific sub-questions:

- **SRQ1:** How do practitioners currently interpret and operationalise key regulatory-aligned data quality dimensions in their day-to-day ML workflows, and what factors shape these practices?
- **SRQ2:** What tools, methods, and infrastructures do practitioners use to manage regulatory-aligned data quality today, and what additional capabilities do they perceive as necessary to meet regulatory obligations?
- **SRQ3:** How do collaboration patterns between technical teams and legal or compliance teams influence the implementation of regulatory-aligned data quality?

To answer these questions, we conducted semi-structured interviews with fourteen EU-based practitioners working on ML systems that process personal data or operate in regulated domains. Participants covered a range of roles, including data engineering, data science, ML engineering, and compliance. The interviews were structured around nine data quality dimensions including accuracy, traceability, and bias mitigation, among others (see Figure 1). Participants were asked to reflect on their experience with data quality issues that arise from legal compliance requirements. We also introduced three scenario-based vignettes inspired by the GDPR

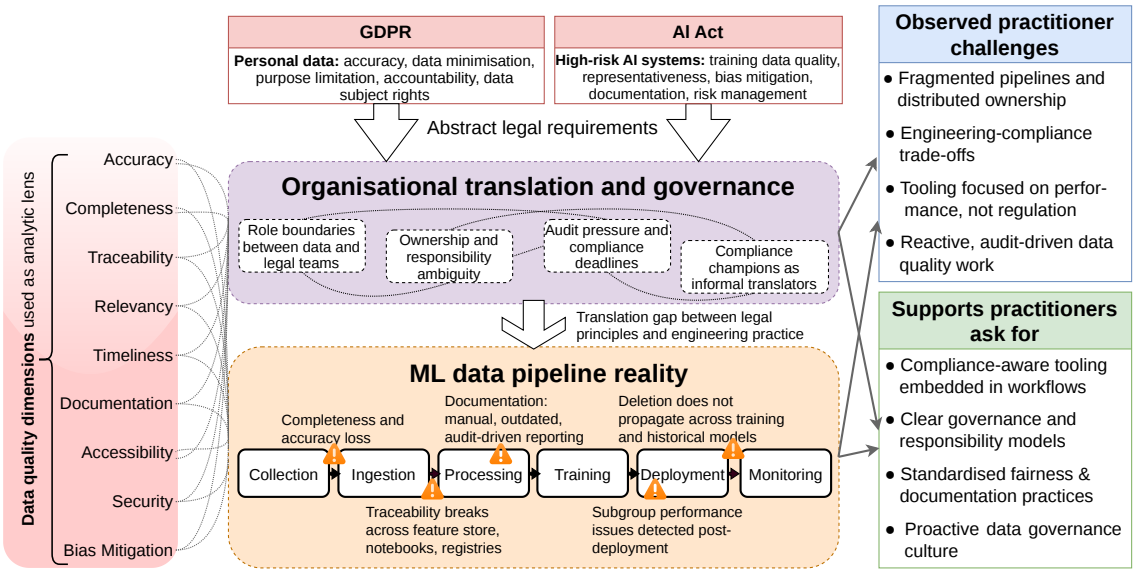

Fig. 1. Regulatory-Aligned Data Quality in Practice - Challenges and Desiderata

and AI Act requirements (see Appendix C). This design encouraged participants to reflect on concrete practices rather than abstract principles.

This paper makes three main contributions. First, it provides an empirical account of how practitioners in the EU interpret and enact regulatory-aligned data quality across ML pipelines. Second, it identifies cross-cutting challenges that shape this work, including the translation of legal principles into engineering tasks, fragmentation of data pipelines, limitations of existing tools, and gaps in collaboration and responsibility sharing between technical and legal teams. Third, it distils practitioner-informed directions for future research on compliance-aware data quality, including opportunities for new forms of technical infrastructure, governance structures, and operational guidance. Figure 1 provides an overview of the study design and analysis outcomes.

2 Literature review

This study sits at the intersection of data quality and EU regulations. Prior work has developed rich technical methods, governance frameworks, and legal analyses. Yet these strands often evolve separately. In this section, we review major literature from technical, regulatory, and empirical perspectives, and identify gaps that motivate our practitioner-focused study of regulatory-aligned data quality in practice.

2.1 Data quality & ML: the technical perspective

Data quality has traditionally been studied in data management through dimensions such as accuracy, completeness, timeliness, consistency, relevance, and accessibility [6, 79]. In ML, data quality is increasingly recognised as a primary determinant of model performance and robustness. Recent work emphasises data-centric AI, arguing that systematic data improvement such as curation, coverage, and label integrity often yields larger gains than model complexity [65, 83]. Studies highlight how label noise [52], dataset shift [60], class imbalance [40], and representativeness [14] affect model behaviour and generalisation.

Quality assurance practices in ML pipelines reflect these concerns. Teams commonly employ schema validation, rule-based checks, and statistical tests to detect anomalies and drift [10, 69]. Feature stores and data versioning systems aim to stabilise training and inference by controlling data dependencies [7, 55]. Experiment tracking frameworks record configurations and metrics to support reproducibility [82]. Tools such as Great Expectations, Deequ, and similar frameworks allow teams to declare and test data quality constraints in code [30, 68]. These practices are now standard in ML operations. However, they primarily prioritise improvements to predictive performance or system reliability and do not always address the requirements associated with legal compliance.

Alongside quality checks, technical governance mechanisms have emerged. Data catalogs, lineage systems, and provenance frameworks track data flows and transformations across pipelines [11, 15, 34]. Metadata management describes datasets, features, and models [67]. Documentation artefacts such as datasheets for datasets and model cards aim to improve transparency and reviewability [25, 50]. Systems such as Apache Atlas, Amundsen, DataHub, and commercial data catalogues aim to provide end-to-end visibility of data assets [2, 31, 45]. MLOps platforms like MLflow, Kubeflow, and cloud-native solutions support experiment tracking, model versioning, and deployment [42, 82].

Yet important gaps remain. Data quality is conceptualised as properties that can be instrumented and monitored within the technical stack. Such infrastructure captures how data moves and how models are trained. It rarely encodes legal requirements such as purpose limitation, lawful basis, consent, or retention constraints. Deletion propagation across derived features, caches, and historical training sets is also weak [16, 27, 38]. Fairness and subgroup evaluation often rely on ad-hoc scripts that are hard to reproduce [35, 49, 61]. Documentation is frequently manual and drifts from the living pipeline, especially in fast-evolving systems [28, 33, 37]. Fragmentation across tools creates blind spots in end-to-end oversight. The gap between technical quality and regulatory quality is therefore built into many current approaches. Our findings corroborate this body of work and provide empirical evidence of how such gaps surface in practitioners' day-to-day compliance work.

2.2 Regulatory perspectives on data quality

EU regulations impose explicit demands on data quality, though they do so at a high level of abstraction. The GDPR regulates the processing of datasets containing personal data through principles of accuracy, data minimisation, purpose limitation, storage limitation, integrity and confidentiality, and accountability [22]. These principles directly affect how personal data is collected, processed, retained, and documented. Accuracy links to record correctness and rectification duties. Minimisation and purpose limitation constrain feature selection and data reuse. Storage limitation determines retention and deletion. Integrity and confidentiality drive access control and protection. Accountability requires evidence and documentation that connect choices to legal obligations and demonstrate compliance with those obligations. Data subject rights further introduce technical demands for traceability, rectification, erasure, and portability [74]. These rights turn technical lineage into compliance evidence. They also expose weak spots in pipelines that were not designed to support such obligations.

The AI Act extends these requirements for high-risk AI systems. Art. 10 AI Act (high-risk AI data governance) becomes applicable in August 2026 (though this may be postponed pending the adoption of the Digital Omnibus on AI). Although these provisions are not yet applicable, they already shape emerging discussions on data quality practices in ML systems. The AI Act mandates data governance and management practices, representative and bias-mitigated training data, risk management, documentation, and post-market monitoring [23]. Legal scholarship has analysed the normative goals of these provisions and their implications for AI accountability [46, 75]. However, these duties are principle-based and technology-neutral [24]. The regulations specify targets but leave implementation choices open.

Applying legal requirements to ML projects requires case-specific interpretation, which creates a persistent translation problem [8, 32, 66, 75, 77, 81]. Regulations articulate abstract legal requirements, while practitioners

need actions and artefacts. Prior interdisciplinary work highlights the difficulty of operationalising legal norms in technical systems [5, 20, 21, 26, 76]. Studies note that engineers are often left to interpret compliance obligations without sufficient guidance [3, 54, 72, 78]. As our findings later show, ML practitioners expressed a clear need to know which pipeline steps, artefacts, and checks provide sufficient evidence. They need shared vocabulary across legal and technical teams.

2.3 Empirical studies on practitioner experience and organisational dynamics

A growing body of empirical research examines how ML practitioners engage with responsible AI practices. Interview and survey studies document challenges related to fairness, explainability, and accountability [17, 35, 48, 63]. These studies show that responsibility is often diffused across teams, incentives are misaligned, and ethical practices are applied late in development cycles.

Research on governance and provenance adoption reveals similar patterns. Practitioners value traceability but struggle to maintain it across organisational boundaries and heterogeneous toolchains [35, 65]. Accountability and normative concerns are rarely embedded into everyday development workflows, instead surfacing reactively after key technical decisions have already been made [51, 57]. In practice, responsibility for governance often falls to informal compliance champions who translate between legal expectations and technical workflows, but these roles are typically under-resourced and lack formal authority [18, 49].

Existing empirical work tends to focus on model-level concerns such as fairness or explainability [48, 63]. Data quality is rarely examined as a cross-cutting concern that spans ingestion, transformation, training, deployment, and monitoring [9]. EU-specific regulatory contexts are also underrepresented. As a result, little is known about how practitioners operationalise data quality dimensions in response to GDPR and AI Act demands. This study directly addresses this gap.

2.4 Research gap

The technical literature offers strong methods for validation, monitoring, versioning, and lineage, yet largely orientated toward performance and operational stability. Regulatory and social-legal work articulates data protection requirements at a high level but provides limited guidance on how these requirements should be operationalised in concrete ML projects. Empirical studies reveal organisational and cultural barriers. Taken together, these strands address related concerns but remain weakly connected.

This study contributes by examining how EU-based practitioners working with personal data or high-risk ML systems navigate this space in practice. Through qualitative interviews, we investigate how they interpret key data quality dimensions, which challenges they encounter when aligning their work with regulatory requirements, and what support they feel is missing. This perspective complements technical and legal approaches and provides an empirical basis for designing future tools, processes, and policies.

3 Methodology

We adopt a qualitative, interview-based design, which is widely used in computer science research to investigate practitioners' experiences and trade-offs in real-world ML development contexts [35, 43, 47]. We interviewed fourteen practitioners based in the EU. In this study, we defined practitioners broadly as individuals involved in any stage of developing ML-based products or services. The sample included operational roles (e.g., data collectors and labellers), technical roles (e.g., data engineers, data scientists, and ML engineers), and legal or compliance roles (e.g., compliance officers and legal staff). Table 1 presents an overview of participant characteristics.

Table 1. Overview of interview respondents, including industry domain, organisational size, role, and core responsibilities.

	Industry	Company Size	Role	Primary Responsibility
R1	Finance (banking)	>10,000	Data scientist/data engineer	ML model development and compliance monitoring
R2	Finance (banking)	>10,000	Data scientist	ML model validation for compliance
R3	Finance, telecommunications	>10,000	Data engineer	Data engineering (freelance)
R4	Retail	>10,000	ML engineer	Sales forecasting models
R5	Finance, transportation	>10,000	Data scientist	Fraud prediction modelling
R6	Transportation	500–1,000	Data analyst	HR analytics dashboards
R7	Health	1,000–5,000	Data scientist/researcher	Disease prediction models
R8	Market research	25–100	Data scientist	Interview data analysis with LLMs
R9	GenAI, transportation	>10,000	Data science manager	Data science oversight and compliance
R10	Finance, geography	1,000–5,000	Data scientist/data collector	Property price forecasting
R11	Audit (financial services)	1,000–5,000	Legal and compliance	Compliance documentation and reporting
R12	Audit (healthcare)	1,000–5,000	Legal and compliance	Privacy compliance
R13	Audit (transportation)	1,000–5,000	Legal and compliance	Risk management consulting
R14	Audit (multiple industries)	1,000–5,000	Legal and compliance	Coordinating compliance reviews

3.1 Interview design

The semi-structured interview protocol contained four main parts (see Appendix B). First, we asked about participants' working background and project setup, including any existing data quality standards. Second, we discussed a concrete case of a data quality issue that became visible due to compliance risk. Third, we introduced a set of short vignettes [39] to explore how participants would act in specific regulatory or data quality scenarios. Finally, for respondents with substantial compliance experience, we asked additional questions about collaboration with legal teams, perceived challenges, trade-offs, and motivations. The protocol was adapted to each participant's role and expertise.

Designing interviews on compliance-related topics requires care. Respondents might hesitate to speak openly due to social desirability concerns or perceived organisational risk. To reduce this pressure, we used several strategies. We explained clearly that all identifying details about individuals, organisations, products, and services would be anonymised. Before scheduling formal interviews, we held informal conversations with most participants to build trust and reduce concerns. During the interviews, we used neutral and technical language to minimise discomfort. We emphasised that the study focuses on understanding challenges and support needs, not on assessing compliance performance. We avoided any phrasing that could imply blame or wrongdoing (examples in Appendix B). Participants were reminded that they could pause or stop the interview or recording at any time. The complete interview guide is available in Appendix B.

As part of this careful design, we used vignettes as an abstract, scenario-based elicitation technique. The vignettes described hypothetical situations unrelated to participants' own organisations, which help reduce defensiveness while prompting reflection on realistic practices and trade-offs. Prior work shows that vignettes are effective for eliciting tacit reasoning and feasibility constraints in ethically or legally sensitive domains [4, 39]. The vignettes were designed to reflect recurring tensions between engineering practice and specific regulatory provisions [5, 36, 46, 73–75], including bias and subgroup performance (GDPR special categories of personal data, Art. 22(4); AI Act data governance, Art. 10(2), (3), and (5)), unclear dataset provenance (GDPR lawful basis, Art. 6; AI Act data governance, Art. 10), and audit preparation (GDPR accountability, Art. 5(2); AI Act documentation,

Table 2. Summary of interview vignettes and their regulatory grounding.

Vignette	Scenario	Regulatory requirements
V1: Subgroup performance	A deployed model performs significantly worse for a subgroup, e.g., by demographic characteristic	Art. 22(4) GDPR; Art. 10(2), (3), (5) AI Act [46, 73, 75]
V2: Dataset provenance	A third-party dataset with incomplete documentation is pushed for fast adoption by management	Art. 6 GDPR; Art. 10 AI Act [5, 36, 75]
V3: Audit preparation	A regulator requests evidence of data quality, bias assessment, and documentation of data handling	Art. 5(2) GDPR; Art. 18 AI Act [36, 74, 75]

Art. 18). Participants were explicitly encouraged to describe how their teams would realistically respond, including what would not be possible or would require additional support. Table 2 summarises the three vignettes and their regulatory grounding; full scripts are provided in Appendix C.

3.2 Recruitment and Implementation

We recruited EU-based participants through convenience sampling, primarily via professional networks and in-person ML and AI events, including conferences, meet-ups, and seminars. To ensure relevance for both the GDPR and the AI Act, we intentionally included participants whose work involved personal data (as defined by Art. 4(1) GDPR) or high-risk ML systems (as listed in Annex III AI Act [23]). Recruitment and interviewing were conducted iteratively between April and December 2025. Data collection and analysis proceeded in parallel, allowing emerging themes to inform continued recruitment. Interviews continued until thematic saturation was reached. After approximately the tenth interview, no materially new codes emerged, and subsequent interviews primarily reinforced existing themes. The codebook and examples of code development are provided in Appendix D.

Interviews were conducted in person or via video conferencing based on participant preference. Each session lasted approximately one hour. We recorded audio for transcription. All personal and organisational identifiers were removed during transcription, and original audio files were deleted afterwards. All transcripts and notes were stored securely on local encrypted devices.

3.3 Analysis

We followed the iterative coding approach described by Deterding and Waters [19], combining inductive and deductive coding strategies. We began with deductive codes derived from established data quality dimensions, relevant EU regulations (the GDPR and the AI Act), and topics from our research questions. We applied these codes systematically to transcripts and field notes. Simultaneously, we remained open to emergent themes not anticipated by our initial framework. We then conducted a second round of analytic coding to identify patterns, similarities, and differences across participants. Emerging themes were refined through discussions with researchers specialising in ML, AI, and law. This iterative process enabled us to identify patterns of meaning that informed the final results.

3.4 Limitations

Care was taken to minimise bias throughout the study, including during interview design, recruitment, and analysis. However, several limitations should be acknowledged.

First, the recruitment process introduces sampling constraints. We aimed to include diverse domains, roles, countries, and organisational sizes, but the overall number of participants remains limited. The sample also skews toward male and larger companies. This reflects broader patterns in the AI practitioner workforce, which remains disproportionately male and concentrated in large companies [1]. A broader sample collected over a longer period

would improve representativeness. That said, the sample size aligns with qualitative research norms, for example, the average sample size in ACM CHI studies is twelve [12]. The sample also contains a role-based clustering effect: four participants (R11–R14) hold legal or compliance roles within audit organisations. Although they work across multiple client industries including finance, healthcare, and transportation, among others, their perspectives amplify compliance-related concerns relative to engineering roles. In addition, the field of ML and data engineering evolves quickly. Participants' accounts of specific tools may become outdated. Nonetheless, many of the underlying needs they described, such as clearer governance and support for regulatory alignment, are likely to persist.

Second, the study relies on an interview approach, which captures perceptions rather than direct observation of organisational behaviour. Participants' accounts conflate factual descriptions with assumptions or incomplete information shaped by their experience, knowledge level, and organisational context. Self-reporting and perception bias are therefore possible. We also did not have direct access to internal systems or organisational workflows, which limits our ability to verify or contextualise their statements. Field studies or ethnographic observations would complement this approach and provide richer evidence of how regulatory-aligned data quality is enacted in practice.

Overall, this study is an initial step toward understanding the challenges and needs of data practitioners working with regulatory requirements. It offers a snapshot of current practice and highlights areas for future research.

4 Interview Results

Across both technical and compliance roles, participants showed a broad awareness of regulatory requirements, especially those related to the GDPR. This is expected, as respondents were selected for having at least some experience with compliance demands. Many described the GDPR as a routine part of their work, while awareness of the AI Act was far less consistent. Some differences emerged across role types. Participants in legal and compliance roles emphasised auditability, documentation, and regulatory interpretation, whereas engineering roles more frequently focused on pipeline fragmentation, tooling limitations, and operational trade-offs. Findings on collaboration gaps and responsibility ambiguity were consistently reported across both groups. Knowledge also varied by industry. Participants from banking or finance reported stronger familiarity with regulatory frameworks than those in less regulated sectors. Practitioners described a mix of motivations behind their data quality and compliance practices. The majority of participants were driven by external pressures such as audits, legal obligations, and fear of fines. Participant R6 (data analyst) said, "GDPR fines are real. As a small company, if we get big fines, we are likely to be bankrupt, so we are extra careful." Others were motivated by internal values, including organisational culture, leadership expectations, and personal responsibility. As participant R5 (data scientist) put it,

> "As a customer, I don't want my data to be carelessly treated. So I become more cautious when processing users' personal data."

These motivations interacted with several factors that influence priorities, including the sensitivity of the data being processed, the organisation's maturity in data governance, the availability of tools, and whether the industry is heavily regulated. Although awareness of compliance was widespread, participants described significant uncertainty about how to translate legal principles into daily engineering tasks and how to navigate the resulting trade-offs. The following themes present the real-world challenges they reported and the kinds of support they identified as necessary.

4.1 Real-world challenges in ensuring regulatory-aligned data quality

Fragmented data pipelines and technical difficulty hindering consistent data quality practices. Participants often emphasised that many of their regulatory-aligned data quality challenges stem directly from the fragmented way data flows through modern ML pipelines. Many described that when data moves across organisational boundaries, the teams responsible for upstream ingestion and engineering often lack the business context or domain knowledge held by the downstream modelling teams. This disconnect frequently undermines data *accuracy* and *completeness*. As practitioner R7 (data scientist/researcher) explained, upstream teams "don't always know what actually matters for the model," making it difficult for them to define appropriate validation constraints or recognise when important attributes are missing or incorrect. Another participant R9 (data science manager) summarised: "The data starts in the upstream team, and by the time it reaches us, no one really knows who validated what." When quality checks are misaligned, downstream teams inherit data they cannot fully trust but are still responsible for using it in regulated ML systems.

This fragmentation also makes *traceability* fragile. Several practitioners described that when data enters complex ML pipelines, they lose visibility into earlier transformations. This makes traceability "partial at best" as data passes through ingestion services, transformation pipelines, feature stores, and deployed models. Even when local logging practices exist, they rarely connect the full path from raw source to model input. This makes it difficult to demonstrate compliance with requirements for auditability or provenance. Participant R2 (data scientist) joked that if they had to recreate the full lifecycle of a dataset from source system to final model, they would have to "chase down five different teams and seven different tools." *Documentation* practices mirror this fragmentation as well. Participants described documentation as "telling different stories" across Git, notebooks, and internal wikis, with no mechanism to keep these sources aligned as data updates through the pipeline. Practitioners noted that these gaps become especially visible during audit reviews, when teams "realise everyone's documentation tells a slightly different story," as noted by participant R10 (data scientist/data collector). Fragmentation also affects *accessibility* and *security*, since data permissions are typically managed by infrastructure teams rather than ML teams, resulting in both over-permissive access in some areas and bottlenecks in others.

The challenges become especially acute when regulations require personal data deletion. Participants repeatedly highlighted that deletion requests cannot be reliably propagated across an ML pipeline. Even when a record can be removed from a primary table, participants explained that traces of it persist throughout the ML lifecycle in derived features, historical datasets, and intermediate extracts. As one engineer, R5, put it,

> "If a user asks for deletion, how do we remove them from derived features or historical model training data? Retraining becomes expensive or infeasible."

Across organisations, even motivated teams struggle to uphold data quality and legal compliance requirements when no one holds responsibility for the full lifecycle of the data. The result is pockets of good practice, but never the consistent, end-to-end governance that regulatory regimes assume [65].

Trade-offs between engineering goals and regulatory data quality requirements. Participants described frequent tensions between what makes a model perform well and what makes it compliant. Through the vignettes, practitioners were prompted to reflect on how technical decisions collide with *fairness, accountability*, and *documentation* requirements in real settings. They helped surface how practitioners would approach these trade-offs in realistic scenarios.

The vignette on **bias and subgroup performance** revealed that addressing bias after deployment was one of the most difficult challenges. Several explained that models often perform well on aggregate metrics but hide significant drops for specific groups. They noted that fixing these gaps required more than retraining. Teams needed additional data collection, features redesign, or specialised evaluation pipelines. Yet these steps were hard to justify under strict delivery timelines. One engineer, R1, said,

> "We want to fix it, but the business still wants the model live. It becomes a negotiation rather than a quality decision."

Participants also noted that fairness fixes can reduce model accuracy for the overall population. Several noted that fairness fixes rarely had clear acceptance criteria. This forced teams to weigh performance losses against the need to mitigate bias. The trade-off was rarely straightforward.

The vignette about **unclear dataset provenance** triggered strong reactions. Practitioners said they often encounter promising third-party datasets with incomplete *documentation*. Many wanted to use the data because it promised performance gains, but they also recognised the regulatory risks. Several respondents in audit and compliance roles said they would push back against management pressure. They would block the dataset until lawful basis and collection practices were clarified. Others admitted that in fast-paced environments, pressure to deliver sometimes wins. R4 (ML engineer) said they might run experiments with the dataset "just to see if it helps", even though they knew it created compliance debt and that it did not meet the GDPR standards for lawful basis or the AI Act governance. Data engineer R3 said, "The dataset might help the model, but if we cannot trust the source, it becomes a time bomb." This scenario revealed a deeper pattern. Performance gains are highly valued, but provenance checks are slow and uncertain. Teams are caught between innovation incentives and accountability requirements.

The vignette on **audit preparation** highlighted a trade-off between delivery speed and the ongoing investment required to maintain audit-ready evidence. Participants said they could provide some evidence, such as accuracy checks, basic lineage, or examples of bias tests. Yet many openly admitted that they would struggle to produce the full documentation required under the GDPR accountability or the AI Act documentation keeping. They had implicitly chosen delivery speed over documentation investment, leaving no coherent, end-to-end narrative linking data collection to model deployment. Participant R3 (data engineer) said they could only show "what we tested last month, not what happened in every step of the pipeline." Teams know how to monitor, validate, and measure performance, but they lack systems that turn these activities into audit-ready compliance evidence.

Across vignettes, a clear pattern emerged. Practitioners understand regulatory expectations, but they operate in environments where engineering goals, time pressure, and organisational structures shape what is feasible. These trade-offs are not signs of neglect. They reflect the reality that compliance and engineering are driven by different incentives and timelines. Participants described a constant balancing act, where they try to meet legal standards while keeping systems functional and competitive.

Lack of tooling that supports regulatory-aligned data quality. Participants described a growing mismatch between the tools used for day-to-day data quality management and the requirements set by regulatory frameworks. Their teams rely on a mix of commercial platforms, open-source libraries, and custom-built systems, but most of these tools target functional or performance-orientated quality. They rarely support the accountability, explainability, or auditability required under EU regulations. As a result, practitioners must fill the gaps with manual work, especially when preparing compliance documentation or risk assessments.

Participants explained that commercial tools were often too rigid to reflect domain-specific constraints, while open-source solutions offered useful components but few compliance-focused features. Many participants said they had built internal validators or monitoring layers to compensate. As engineer R3 put it, their custom tool became "the glue we needed to make sense of our data, because nothing off-the-shelf understood our domain or what compliance meant for us." Others explained that they created internal systems simply to match the pace of product development, noting that only large organisations have the resources to do this. For smaller companies, keeping up with compliance was described as even more challenging.

The tooling deficit extended into the broader ML infrastructure. MLOps platforms such as MLflow, Azure ML, and SageMaker were valued for versioning and reproducibility, but participants criticised them for lacking features needed for regulatory-aligned data quality, such as capturing purpose, legal basis, or retention rules.

Versioning tools could record *what* changed but not *why*, and lineage systems showed technical transformations but not their compliance implications. Several participants noted that small changes between model versions, such as adding a feature from a new data source or using a longer history of data, could introduce different legal bases, retention rules, or fairness risks. Their tools could record the technical version change but could not reveal these regulatory consequences. When their teams needed to justify why certain fields were included in a model, they often had to reconstruct the story manually. Participant R2 (data scientist) described this process as "detective work we shouldn't have to do".

These gaps were also evident in participants' *methods* and *documentation* practices. Teams used a combination of statistical validation, distribution monitoring, bias testing, and human-in-the-loop review to assess data quality, but noted that these methods were designed around model performance rather than regulatory needs. Fairness assessments commonly relied on ad-hoc scripts, not standardised pipelines, and no practitioner reported having unified tooling for *bias* evaluation. *Documentation* workflows were similarly fragmented. While practitioners referenced datasheets, model cards, and metadata tables, these materials were usually created after the fact, disconnected from pipeline execution, and quickly outdated. Participant R2 (data scientist) in the financial industry described spending weeks each year producing a 200-page compliance report that translated their technical workflow into plain language. Even with the introduction of generative AI tools or internal chatbots, participants (R5, data scientist and R6, data analyst) said these assistants could summarise code or metadata but could not keep documentation automatically synchronised with real ML pipelines, forcing teams back to manual editing.

Altogether, participants' experiences revealed a tooling ecosystem orientated towards engineering performance but not for regulatory alignment. Existing tools fall short of supporting the consistency, traceability, explainability, and documentation required for compliance. This leaves practitioners dependent on manual processes, ad-hoc scripts, domain intuition, and cross-team negotiation, which are difficult to sustain as regulatory scrutiny increases.

Collaboration and responsibility gaps between data and legal teams. Participants repeatedly emphasised that many of their challenges were not purely technical but stemmed from organisational and communication gaps between data teams and legal or compliance teams. They described considerable ambiguity in translating high-level regulatory principles from the GDPR or the AI Act into concrete, operational data quality practices. Responsibility for specific tasks was often unclear. Practitioner R6 (data analyst) captured this confusion by noting,

> "We thought legal was supposed to review these procedures. Legal thought we were controlling it."

These blurred boundaries created uncertainty and delays, especially when a system neared deployment or an audit approached. Collaboration practices reflected these structural tensions. Participants explained that cooperation between data and legal teams was usually informal and reactive. Both sides expressed a desire to be involved earlier in the process, rather than being consulted only at the end. When compliance entered late, technical teams struggled to adjust pipeline design, feature sets, or documentation under time pressure. As practitioner R6 (data analyst) put it, "We hope we can be involved earlier in the compliance procedures." Legal teams expressed similar frustrations, noting that earlier engagement would help them provide more meaningful guidance. In some organisations, technical teams relied on internal compliance champions to bridge this gap, but these individuals carried significant workloads and spent considerable time translating regulatory guidance into actionable engineering steps.

These accounts show that collaboration and responsibility boundaries between data and legal teams remain loosely defined in many organisations. Without shared language, regular communication, or clearer ownership, regulatory-aligned data quality relies heavily on cross-team negotiation and individual initiative. The result is a compliance process driven by ad-hoc negotiation instead of a stable, integrated workflow.

Data quality work is largely invisible, unrewarded, under-resourced, and reactive. Participants described data quality work as essential but often receiving limited attention, especially when it carried regulatory implications. Many described it as invisible labour that is necessary but rarely planned, rewarded, or adequately resourced. As a result, regulatory-aligned quality practices tend to be surfaced and triggered by audits or incidents, rather than embedded proactively into ML workflow.

Several practitioners noted that their teams tried to reconstruct lineage only after something failed or when a regulator requested evidence. Engineer R3 said their team only created lineage diagrams "when something goes wrong," and even then, the process was manual and incomplete. Participants also noted compromises in security and access control. Under delivery pressure, teams sometimes granted broad access to speed up development, even when this conflicted with internal policy. Review cycles and regular risk assessments were also rare. Some respondents said compliance was viewed as a one-time hurdle. Once a legal team approved the initial design, teams assumed they were "free to do whatever we want" until the next major review (R11, compliance officer). This mindset made it difficult to maintain regulatory-aligned data quality as models evolved or as new data sources were introduced.

Across interviews, participants shared a sense that regulatory-aligned data quality work becomes visible only when problems emerge. They said it required time, coordination, and efforts that were rarely recognised in performance evaluations or project timelines. Without dedicated time, resources, and recognition, regulatory-aligned data quality is often addressed reactively rather than as a routine part of ML practice.

4.2 Support practitioners need to manage regulatory-aligned data quality

Technical support needed. Participants said they needed tools that could connect engineering work with regulatory requirements. Many current systems could validate data or track model versions, but they did not help teams understand legal implications. Practitioners asked for compliance-aware tooling that fits directly into ML workflows. They wanted lineage systems that record not only how data moves but also why it is allowed to move, including purpose, retention rules, and data source conditions. Several also mentioned the need for deletion and retention mechanisms that work across pipelines, rather than only at the storage layer. Participants said this would reduce the manual "detective work" they now perform when responding to audits or deletion requests. Fairness and bias tools were another gap. People wanted standardised, repeatable fairness assessments, not ad-hoc scripts. Documentation tools that stay in sync with pipelines were also seen as essential. They called for platforms that surface risks early and give clear signals when a dataset, feature, or model version requires legal review. Participant R1 (data scientist/data engineer) highlighted this,

> "We want tools that tell us the compliance story, not just the technical one."

Organisational and process-level support needed. Practitioners emphasised that technical tools alone were not enough. They needed clearer governance structures and stable processes to support regulatory-aligned data quality management. Participants called for explicit roles that assign responsibility, along with regular review cycles. They also wanted shared procedures for dataset approval, fairness evaluation, and audit preparation. Some said they wished for standard templates or checklists that would guide them through regulated steps, especially when preparing documentation or evaluating new datasets. They wanted simple explanations of what regulators expect and how these expectations connect to daily engineering tasks. Some suggested cross-functional working groups that include data teams, legal teams, and product owners. This would help teams understand what good compliance looks like and when to involve legal specialists.

Cultural support and mindset shifts. Participants said that meeting regulatory-aligned data quality also needs a cultural shift. Many described a mindset where data quality and compliance are only done when required rather than ongoing practices. They wanted leadership to recognise that quality work takes time and should be included in project planning. They hoped for a culture where teams address quality proactively, rather than

reacting only when audits or incidents occur. Others noted that organisations should reward work that prevents issues, not only work that delivers new features. Practitioners also emphasised the need for shared language between technical and legal staff. A more collaborative culture, supported by early involvement from all teams, was seen as essential for sustainable and trustworthy ML systems.

5 Discussion and Recommendations

This study examined how practitioners understand and operationalise regulatory-aligned data quality in their daily work. The findings show that practitioners are aware of data quality and compliance demands, yet struggle to translate abstract regulatory rules into concrete engineering practice. Together, these insights help reframe data quality as a socio-technical process shaped by regulation, governance, and engineering constraints. In this section, we discuss the broader meaning of these findings and outline implications for technical practice, organisational structures, and future research.

Reinterpreting data quality as a socio-technical practice. Data and ML practitioners often approached data quality from an engineering perspective, focusing on performance, schema validation, distribution checks, or pipeline stability. Regulatory frameworks, however, frame data quality in terms of accountability, rights protection, and demonstrable compliance. Our study shows that these differences create uncertainty about how to translate compliance requirements in everyday work. When engineers validate data, they test for missing values and distributional anomalies. These checks do not test for population representativeness, as will be required by Art. 10 AI Act, nor for purpose limitation, as required by Art. 5(1)(b) GDPR. Lineage tools record technical transformations but not the legal justifications. This mirrors the pattern identified for fairness: technical abstractions can strip away the contextual nuance that gives legal concepts their regulatory force [71, 81]. Our contribution is to show that this dynamic extends beyond fairness to all major data quality dimensions, including accuracy, completeness, traceability, and relevance. Lifecycle-orientated documentation offers a structural response [37], but our evidence suggests it remains largely aspirational in practice. Practitioners cannot construct the end-to-end narrative it requires, because regulatory constraints must persist across transformations and retraining rather than drift as static artefacts. Regulatory-aligned data quality is therefore not a static property of data. Instead, it is a socio-technical practice, in which technical infrastructures and regulatory expectations must be co-designed.

The regulatory-technical translation gap. A central insight of this study is that the translation gap between legal requirements and engineering practice is not a lack of willingness. Instead, the problem is a lack of structures that help teams interpret what those requirements mean for specific pipeline steps, artefacts, and decision points. Prior work has shown that normative requirements are typically handled without adequate guidance and pushed to late in development cycles [3, 5, 20, 35, 72, 78]. Our empirical data suggests that practitioners cannot operationalise a requirement to ensure "appropriate and representative" training data without guidance on what counts as appropriate, for which risk classification, and under what evidentiary standard. When those structures are missing, compliance work is deferred.

More importantly, the gap compounds over time as traceability, documentation, and deletion obligations degrade across evolving data pipelines. Participants described knowingly accumulating “compliance debt” [41]: running experiments with datasets of uncertain legal basis, skipping provenance checks, and granting broad access to accelerate development. What our study adds to this concept is an account of how compliance debt accrues across multiple dimensions simultaneously in ML data pipelines. Traceability, documentation, deletion obligations, and fairness assessment all degrade in parallel as systems evolve and teams turn over.

The practical implication follows directly. Interpretive guidance would help, but it is not sufficient. This points toward a compliance by design principle, analogous to privacy by design [13], where legal requirements are treated as first-class design constraints from the outset of ML pipeline development, rather than addressed through

audit-triggered rework. Practitioners need mechanisms that make the compliance implications of pipeline changes visible at the moment those changes are made, before debt accumulates and before audits force a reckoning.

Implications for technical infrastructure and ML practice. The findings highlight the need for technical solutions that embed compliance into ML workflows. First, existing tools support validation, monitoring, and versioning [69, 82], but they rarely capture the legal metadata required to trace why data was lawfully included, what retention obligation applies, or which downstream components are affected when a legal basis changes. This gap creates significant manual work during audit requests. A compliance-aware infrastructure should associate legal metadata with data assets and model versions, and identify compliance implications automatically when versioned changes are made. Lineage tools must extend beyond technical dependency graphs to support bidirectional reasoning: from model outputs back to source data, and from regulatory obligations forward to affected pipeline components. Similarly, organisations need deletion-aware systems that can propagate removal requests across derived features and historical datasets that often contain residual traces of personal data. Participants asked for lineage tools that reflect both technical and legal dimensions. Second, documentation should be generated during pipeline execution rather than reconstructed after the fact [37, 54]. Third, participants also called for more robust fairness assessment as a standardised pipeline component with reproducible subgroup evaluations, not ad-hoc scripts [9]. These gaps point toward ML infrastructure that treats legal context as a first-class concern rather than an external reporting task.

Implications for organisational structures and governance. Technical improvements alone will not close the compliance gap. Our findings show that organisational structures play a crucial role in regulatory-aligned data quality. Responsibility for quality tasks was often unclear. Data teams assumed that legal teams would provide detailed rules, while legal teams expected engineers to design compliant workflows. This gap created delays, uncertainty, and compliance risks. Early involvement of both sides was seen as essential, yet often missing. This stands in tension with Art. 25(1) GDPR's requirement for data protection by design and by default, which presupposes early consideration of compliance. These dynamics align with findings on collaboration breakdowns in fairness work [17, 18, 35, 49], and our study shows the same structural barriers govern legal compliance. In this domain, evidentiary failures cannot be retroactively fixed, making clearer governance structures all the more critical. Concretely, organisations need defined ownership over compliance-relevant tasks such as documentation, fairness evaluation, and dataset approval. Compliance champion roles need formal authority and dedicated time, not merely informal expectations. Cross-functional review cycles should be scheduled into project timelines, not triggered by audits. Shared internal guidelines can also help surface compliance issues earlier and reduce last-minute rework. Incentives such as documentation currency, deletion compliance rate, and fairness assessment coverage would make compliance work organisationally visible and reward it on par with model performance. Without these structural reforms, compliance remains inconsistent and often reactive.

Promoting a culture of regulatory-aligned data quality. Nonetheless, the key challenge is not simply the absence of procedures. Rather, regulatory-aligned data quality work remains largely invisible, under-resourced, and weakly rewarded in everyday ML practice. This aligns with patterns reported in responsible AI and fairness literature: normative work is deprioritised in favour of delivery speed and surviving through individual initiative rather than being embedded in organisational structures [51, 63, 65]. Our findings extend this to regulatory-aligned data quality specifically: practitioners described this work as something that surfaces mainly when legal questions arise.

Rather than adding more processes, organisations should make this work recognised and valued. This means data quality and compliance should not be treated as separate or competing concerns. Regulatory-aligned data quality offers a way to integrate technical data quality work with legal accountability in one operational frame. Concretely, audit readiness should be expected as a normal state, and data quality and compliance contributions should be reflected in key performance indicators (KPIs), rather than treated as extra work. Reducing the effort

required to maintain documentation, traceability, and checks can further support this shift by making such practices easier to sustain in routine workflows. This also points to a broader agenda for future work: developing ways to evaluate and compare how organisations recognise, incentivise, and sustain regulatory-aligned data quality in practice.

6 Conclusion

This paper provides an empirical view of regulatory-aligned data quality as it is experienced by practitioners working with ML systems in the EU. Rather than treating data quality as a purely technical attribute, this paper shows how regulatory expectations reshape everyday data work across pipelines, roles, and organisations.

In addressing **SRQ1**, the study shows how practitioners interpret and operationalise key regulatory-aligned data quality dimensions in practice, and how these interpretations are shaped by regulatory demands. Looking more closely at day-to-day workflows, the findings also address **SRQ2**. We show how existing tools, methods, and infrastructures support performance-orientated notions of data quality. Support for regulatory obligations such as documentation, traceability, and lifecycle governance remains limited. At the same time, practitioners' accounts highlight how these limitations are often managed through manual workarounds and ad-hoc practices. Finally, in relation to **SRQ3**, the study illustrates how collaboration patterns between technical teams and legal or compliance teams influence whether regulatory-aligned data quality can be implemented effectively, and where gaps in responsibility, communication, and timing persist.

By foregrounding practitioner perspectives, this study brings to light the often unseen coordination, interpretation, and trade-off work required to align ML systems with evolving regulatory frameworks. These insights motivate the need for future research that bridges legal intent and technical implementation, and that supports practitioners through compliance-aware infrastructure, clearer governance models, and sustained cross-disciplinary collaboration.

Endmatter

Generative AI Statement

The authors used ChatGPT for grammar checking and stylistic editing.

Ethical Considerations

This study was conducted in accordance with the ACM Code of Ethics and Professional Conduct. The study protocol received approval from the authors' institutional ethics committee. Participation was voluntary and based on informed consent. Given the sensitivity of compliance-related topics, all personal and organisational identifiers were anonymised.

Author Contributions

Yichun Wang designed the study, conducted the interviews, and drafted the paper under the supervision of the other authors. Kristina Irion contributed to the study design, legal analysis, and reviewing and provided supervision on the legal aspects of the work. Paul Groth contributed to the study design and reviewing and provided supervision throughout the project. Hazar Harmouch managed the collaboration and contributed to the study design, reviewing and refining the draft.

Acknowledgments

The authors thank all interview participants for their time, candour, and willingness to share their experiences with regulatory-aligned data quality in practice. Their insights are the foundation of this work. The authors also thank the anonymous reviewers for their constructive feedback and suggestions, which have significantly

improved the paper. This research was performed in the context of the AI4FinTech initiative of the University of Amsterdam.

References

[1] AI Index Steering Committee. 2025. AI Index Report 2025. https://hai.stanford.edu/ai-index/2025-ai-index-report. Accessed: 2025-02-05.

[2] Apache Software Foundation. 2024. Apache Atlas: Data Governance and Metadata framework for Hadoop. https://atlas.apache.org/ Software available from atlas.apache.org.

[3] Masoud Barati, Gagangeet Singh Aujla, Jose Tomas Llanos, Kwabena Adu Duodu, Omer Farooq Rana, Madeline Carr, and Rajiv Ranjan. 2021. Privacy-Aware Cloud Auditing for GDPR Compliance Verification in Online Healthcare. *IEEE Transactions on Industrial Informatics* 18 (2021), 4808–4819. https://api.semanticscholar.org/CorpusID:237726883

[4] Christine Barter and Emma Renold. 2000. 'I wanna tell you a story': exploring the application of vignettes in qualitative research with children and young people. *International journal of social research methodology* 3, 4 (2000), 307–323.

[5] David A. Basin, Søren Debois, and Thomas T. Hildebrandt. 2018. On Purpose and by Necessity: Compliance Under the GDPR. In *Financial Cryptography*. https://api.semanticscholar.org/CorpusID:13748135

[6] Carlo Batini, Monica Scannapieco, et al. 2016. Data and information quality. *Cham, Switzerland: Springer International Publishing* 63 (2016).

[7] Anant Bhardwaj, Souvik Bhattacherjee, Amit Chavan, Amol Deshpande, Aaron J Elmore, Samuel Madden, and Aditya G Parameswaran. 2014. Datahub: Collaborative data science & dataset version management at scale. *arXiv preprint arXiv:1409.0798* (2014).

[8] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on fairness, accountability and transparency*. PMLR, 149–159.

[9] Emily Black, Rakshit Naidu, Rayid Ghani, Kit Rodolfa, Daniel Ho, and Hoda Heidari. 2023. Toward operationalizing pipeline-aware ML fairness: A research agenda for developing practical guidelines and tools. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–11.

[10] Eric Breck, Shanqing Cai, Eric Nielsen, Michael Salib, and D. Sculley. 2017. The ML test score: A rubric for ML production readiness and technical debt reduction. *2017 IEEE International Conference on Big Data (Big Data)* (2017), 1123–1132. https://api.semanticscholar.org/CorpusID:6244440

[11] Peter Buneman, Sanjeev Khanna, and Wang Chiew Tan. 2001. Why and Where: A Characterization of Data Provenance. In *International Conference on Database Theory*. https://api.semanticscholar.org/CorpusID:13791826

[12] Kelly E. Caine. 2016. Local Standards for Sample Size at CHI. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016). https://api.semanticscholar.org/CorpusID:14445824

[13] Ann Cavoukian and Jeff Jonas. 2012. Privacy by design in the age of big data. (2012).

[14] Kyla Chasalow and Karen E. C. Levy. 2021. Representativeness in Statistics, Politics, and Machine Learning. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021). https://api.semanticscholar.org/CorpusID:231573261

[15] James Cheney, Laura Chiticariu, and Wang Chiew Tan. 2009. Provenance in Databases: Why, How, and Where. *Found. Trends Databases* 1 (2009), 379–474. https://api.semanticscholar.org/CorpusID:1778556

[16] Rishav Chourasia, Neil Shah, and R. Shokri. 2022. Forget Unlearning: Towards True Data-Deletion in Machine Learning. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:252917763

[17] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring how machine learning practitioners (try to) use fairness toolkits. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 473–484.

[18] Wesley Hanwen Deng, Nur Yildirim, Monica Chang, Motahhare Eslami, Kenneth Holstein, and Michael Madaio. 2023. Investigating practices and opportunities for cross-functional collaboration around AI fairness in industry practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 705–716.

[19] Nicole M Deterding and Mary C Waters. 2021. Flexible coding of in-depth interviews: A twenty-first-century approach. *Sociological methods & research* 50, 2 (2021), 708–739.

[20] Lilian Edwards and Michael Veale. 2017. Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For. *Duke law and technology review* 16 (2017), 18–84. https://api.semanticscholar.org/CorpusID:59148202

[21] Lilian Edwards and Michael Veale. 2018. Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"? *IEEE Security & Privacy* 16 (2018), 46–54. https://api.semanticscholar.org/CorpusID:4049746

[22] European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). http://data.europa.eu/eli/reg/2016/679/oj Official Journal of the European Union, L 119, 4 May 2016, pp. 1-88.

[23] European Parliament and Council of the European Union. 2024. Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. http://data.europa.eu/eli/reg/2024/1689/oj

[24] Marta Cantero Gamito and Christopher T Marsden. 2024. Artificial intelligence co-regulation? The role of standards in the EU AI Act. *Int. J. Law Inf. Technol.* 32 (2024). https://api.semanticscholar.org/CorpusID:271069058

[25] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé, and Kate Crawford. 2018. Datasheets for datasets. *Commun. ACM* 64 (2018), 86 – 92. https://api.semanticscholar.org/CorpusID:4421027

[26] Marissa Kumar Gerchick, Ro Encarnación, Cole Tanigawa-Lau, Lena Armstrong, Ana Gutiérrez, and Danaé Metaxa. 2025. Auditing the Audits: Lessons for Algorithmic Accountability from Local Law 144's Bias Audits. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 29–44. doi:10.1145/3715275.3732004

[27] Antonio A. Ginart, Melody Y. Guan, Gregory Valiant, and James Y. Zou. 2019. Making AI Forget You: Data Deletion in Machine Learning. *ArXiv* abs/1907.05012 (2019). https://api.semanticscholar.org/CorpusID:195886255

[28] Joan Giner-Miguelez, Abel Gómez, and Jordi Cabot. 2024. Using large language models to enrich the documentation of datasets for machine learning. *arXiv preprint arXiv:2404.15320* (2024).

[29] Michelle Goddard. 2017. The EU General Data Protection Regulation (GDPR): European regulation that has a global impact. *International Journal of Market Research* 59, 6 (2017), 703–705.

[30] Abe Gong, James Campbell, and Great Expectations. [n. d.]. *Great Expectations*. https://github.com/great-expectations/great_expectations

[31] Mark Grover et al. 2019. Amundsen: Lyft's Data Discovery & Metadata Engine. https://github.com/amundsen-io/amundsen Lyft Engineering.

[32] Neil Guha, Christie M Lawrence, Lindsey A Gailmard, Kit T Rodolfa, Faiz Surani, Rishi Bommasani, Inioluwa Deborah Raji, Mariano-Florentino Cuéllar, Colleen Honigsberg, Percy Liang, et al. 2024. AI regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing. *Geo. Wash. L. Rev.* 92 (2024), 1473.

[33] Amy Kathleen Heger, Elizabeth B. Marquis, Mihaela Vorvoreanu, Hanna M. Wallach, and Jennifer Wortman Vaughan. 2022. Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata. *Proceedings of the ACM on Human-Computer Interaction* 6 (2022), 1 – 29. https://api.semanticscholar.org/CorpusID:249431472

[34] Melanie Herschel, Ralf Diestelkämper, and Houssem Ben Lahmar. 2017. A survey on provenance: What for? What form? What from? *The VLDB Journal* 26 (2017), 881 – 906. https://api.semanticscholar.org/CorpusID:32007099

[35] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miroslav Dudík, and Hanna M. Wallach. 2018. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2018). https://api.semanticscholar.org/CorpusID:54895308

[36] Chris Jay Hoofnagle, Bart Van Der Sloot, and Frederik Zuiderveen Borgesius. 2019. The European Union general data protection regulation: what it is and what it means. *Information & Communications Technology Law* 28, 1 (2019), 65–98.

[37] Ben Hutchinson, Andrew Smart, Alex Hanna, Remi Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 560–575.

[38] Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Y. Zou. 2020. Approximate Data Deletion from Machine Learning Models: Algorithms and Evaluations. *ArXiv* abs/2002.10077 (2020). https://api.semanticscholar.org/CorpusID:232033672

[39] Nicholas Jenkins, Michael Bloor, Jan Fischer, Lee Berney, and Joanne Neale. 2010. Putting it in context: the use of vignettes in qualitative interviewing. *Qualitative research* 10, 2 (2010), 175–198.

[40] Justin M. Johnson and Taghi M. Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data* 6 (2019), 1–54. https://api.semanticscholar.org/CorpusID:102354936

[41] Benedikt Kolbeinsson and Arinbjörn Kolbeinsson. 2025. Compliance Debt: A Taxonomy for Deferred Governance in Regulated Systems. In *EurIPS 2025 Workshop on Private AI Governance*. https://openreview.net/forum?id=tdB8sbTuY0

[42] Kubeflow Authors. 2024. Kubeflow: The Machine Learning Toolkit for Kubernetes. https://www.kubeflow.org/ Open-source software project.

[43] Steinar Kvale. 2009. *Interviews: Learning the craft of qualitative research interviewing*. Sage.

[44] Clément Labadie and Christine Legner. 2023. Building data management capabilities to address data protection regulations: Learnings from EU-GDPR. *Journal of Information Technology* 38, 1 (2023), 16–44.

[45] Mars Lan et al. 2019. DataHub: Popularizing Data Discovery at LinkedIn. https://datahubproject.io/ LinkedIn Engineering.

[46] Johann Laux, Sandra Wachter, and Brent Daniel Mittelstadt. 2023. Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance* 18 (2023), 3 – 32. https://api.semanticscholar.org/CorpusID:260498921

[47] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research methods in human-computer interaction*. Morgan Kaufmann.

[48] Michael Madaio, Shivani Kapania, Rida Qadri, Ding Wang, Andrew Zaldivar, Remi Denton, and Lauren Wilcox. 2024. Learning about responsible AI on-the-job: Learning pathways, orientations, and aspirations. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1544–1558.

[49] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020). https://api.semanticscholar.org/CorpusID:210156214

[50] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2018. Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2018). https://api.semanticscholar.org/CorpusID:52946140

[51] Jakob Mökander and Luciano Floridi. 2023. Operationalising AI governance through ethics-based auditing: an industry case study. *AI and Ethics* 3, 2 (2023), 451–468.

[52] Curtis Northcutt, Lu Jiang, and Isaac Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research* 70 (2021), 1373–1411.

[53] Claudio Novelli, Federico Casolari, Antonino Rotolo, Mariarosaria Taddeo, and Luciano Floridi. 2024. Taking AI risks seriously: a new assessment model for the AI Act. *Ai & Society* 39, 5 (2024), 2493–2497.

[54] Victor Ojewale, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. 2025. Towards AI accountability infrastructure: Gaps and opportunities in AI audit tooling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–29.

[55] Laurel Orr, Atindriyo Sanyal, Xiao Ling, Karan Goel, and Megan Leszczynski. 2021. Managing ml pipelines: feature stores and the coming wave of embedding ecosystems. *arXiv preprint arXiv:2108.05053* (2021).

[56] Yaniv Ovadia, Emily Fertig, Jie Jessie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. In *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:174803437

[57] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019). https://api.semanticscholar.org/CorpusID:57759365

[58] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. 2016. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 2233–2241. https://api.semanticscholar.org/CorpusID:16406313

[59] Maria Priestley, Fionntán O'donnell, and Elena Simperl. 2023. A survey of data quality requirements that matter in ML development pipelines. *ACM Journal of Data and Information Quality* 15, 2 (2023), 1–39.

[60] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. 2009. Dataset Shift in Machine Learning. https://api.semanticscholar.org/CorpusID:61294087

[61] Ricardo Trainotti Rabonato and Lilian Berton. 2024. A systematic review of fairness in machine learning. *AI and Ethics* 5 (2024), 1943 – 1954. https://api.semanticscholar.org/CorpusID:272795996

[62] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.

[63] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2020. Where Responsible AI meets Reality. *Proceedings of the ACM on Human-Computer Interaction* 5 (2020), 1 – 23. https://api.semanticscholar.org/CorpusID:219966320

[64] Sandeep Rangineni. 2023. An analysis of data quality requirements for machine learning development pipelines frameworks. *International Journal of Computer Trends and Technology* 71, 9 (2023), 16–27.

[65] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[66] Holli Sargeant and Måns Magnusson. 2025. Formalising Anti-Discrimination Law in Automated Decision Systems. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 181–194.

[67] Pegdwendé N. Sawadogo and Jérôme Darmont. 2020. On data lake architectures and metadata management. *Journal of Intelligent Information Systems* 56 (2020), 97 – 120. https://api.semanticscholar.org/CorpusID:220064078

[68] Sebastian Schelter, Felix Biessmann, Dustin Lange, Tammo Rukat, Phillipp Schmidt, Stephan Seufert, Pierre Brunelle, and Andrey Taptunov. 2019. Unit testing data with deequ. In *Proceedings of the 2019 International Conference on Management of Data*. 1993–1996.

[69] Sebastian Schelter, Dustin Lange, Philipp Schmidt, Meltem Celikel, Felix Biessmann, and Andreas Grafberger. 2018. Automating Large-Scale Data Quality Verification. *Proc. VLDB Endow.* 11 (2018), 1781–1794. https://api.semanticscholar.org/CorpusID:51989857

[70] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. *Advances in neural information processing systems* 28 (2015).

[71] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

[72] Sean Sirur, Jason RC Nurse, and Helena Webb. 2018. Are we there yet? Understanding the challenges faced in complying with the General Data Protection Regulation (GDPR). In *Proceedings of the 2nd international workshop on multimedia privacy and security*. 88–95.

[73] Marvin Van Bekkum and Frederik Zuiderveen Borgesius. 2023. Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception? *Computer Law & Security Review* 48 (2023), 105770.

[74] Michael Veale, Reuben Binns, and Jef Ausloos. 2018. When data protection by design and data subject rights clash. *International Data Privacy Law* 8 (2018), 105–123. https://api.semanticscholar.org/CorpusID:58193452

[75] Michael Veale and Frederik J. Zuiderveen Borgesius. 2021. Demystifying the Draft EU Artificial Intelligence Act — Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International* 22 (2021), 97 – 112. https://api.semanticscholar.org/CorpusID:235765823

[76] Michael Veale and Lilian Edwards. 2018. Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling. *Comput. Law Secur. Rev.* 34 (2018), 398–404. https://api.semanticscholar.org/CorpusID:4797884

[77] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[78] Paul Voigt and Axel von dem Bussche. 2024. The EU General Data Protection Regulation (GDPR): A Practical Guide. *The EU General Data Protection Regulation (GDPR)* (2024). https://api.semanticscholar.org/CorpusID:273725799

[79] Richard Y Wang and Diane M Strong. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems* 12, 4 (1996), 5–33.

[80] Yichun Wang, Kristina Irion, Paul Groth, and Hazar Harmouch. 2026. Machine Learning Practitioners' Views on Data Quality in Light of EU Regulatory Requirements: A European Online Survey. arXiv:2602.06594 [cs.DB] https://arxiv.org/abs/2602.06594

[81] Elizabeth Anne Watkins and Jiahao Chen. 2024. The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 764–775.

[82] Matei A. Zaharia, Andrew Chen, Aaron Davidson, Ali Ghodsi, Sue Ann Hong, Andy Konwinski, Siddharth Murching, Tomas Nykodym, Paul Ogilvie, Mani Parkhe, Fen Xie, and Corey Zumar. 2018. Accelerating the Machine Learning Lifecycle with MLflow. *IEEE Data Eng. Bull.* 41 (2018), 39–45. https://api.semanticscholar.org/CorpusID:83459546

[83] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2025. Data-centric artificial intelligence: A survey. *Comput. Surveys* 57, 5 (2025), 1–42.

A Data Quality Dimensions and Regulatory Mapping

Table 3 maps commonly used data quality dimensions from the data engineering literature to corresponding requirements in the GDPR and the AI Act. These dimensions provide a structured lens for analysing regulatory-aligned data quality and were used to guide the design of the interview protocol, including question prompts and scenario-based vignettes.

B Interview Protocol and Outline

Figure 2 shows the interview outline used to structure the semi-structured interviews.

Note on neutral language strategies. Rather than asking whether participants were "compliant," we asked how their team would "realistically handle" a situation. Questions about responsibility were framed around process and workflow rather than individual accountability. For example, "who is typically involved at this stage" rather than "whose responsibility is this." Regulatory obligations were referred to by their technical counterparts where possible, such as "traceability" and "documentation practices" rather than "legal accountability," to keep the conversation grounded in familiar engineering language.

C Interview Vignettes

The following vignettes were used during the interviews to prompt discussion about regulatory-aligned data quality practices. The scenarios are hypothetical and intentionally abstract, and do not refer to participants' own organisations or systems.

"Some people don't immediately recall compliance examples, so let me give you a short scenario and ask how you'd handle it."

"Please tell us how your team would realistically handle them in your current setting. If something wouldn't be possible, or if you'd need extra support, that's valuable for us to hear too."

"We're not looking for the 'perfect' answer, but rather the feasible choices and trade-offs your team makes in practice."

C.1 Vignette 1: Bias and Subgroup Performance

"Imagine you have already deployed a machine learning model. Overall performance looks fine, but monitoring shows that the model performs significantly worse for a particular subgroup, for example older users or a specific gender group."

"How would your team realistically handle this situation in your current setting?"

"If fixing this would require trade-offs, such as retraining, collecting new data, or accepting lower overall performance, I'm interested in how you would think about those decisions."

C.2 Vignette 2: Unclear Dataset Provenance

"Your team receives a third-party dataset that could improve model accuracy, but the documentation is incomplete. It's unclear how the data were collected, for what purpose, or under what conditions."

"Management would like to use the dataset quickly because of delivery timelines."

"In your current setup, how would your team realistically deal with this?"

"If something wouldn't be possible, or if you would need extra support from legal or other teams, that's also useful for us to understand."

C.3 Vignette 3: Audit Preparation and Documentation

"Now imagine a regulator contacts your organisation."

Table 3. Intersection of data quality dimensions from technical and regulatory perspectives (GDPR and AI Act). These dimensions guided the design of the interview protocol, including question prompts and scenario-based vignettes. Their systematic mapping to regulatory provisions is developed in a companion survey paper [80].

Data Quality Dimensions	Technical Interpretation	Provisions in the GDPR as applicable	Provisions in the AI Act as applicable
Accuracy	Data is correct, reliable, and free from errors.	Requires data accuracy; erase or rectify incorrect personal data – **Art. 5(1)(d)**, **Recital 39** Right to rectification of inaccurate personal data – **Art. 16**	Datasets shall be error-free – **Art. 10(3)** Data sets should be, to the best extent possible, free of errors – **Recital 67**
Traceability	Changes or updates of the data should be tracked.	Security measures including logs – **Art. 32**	Mandatory automated logging and record-keeping for traceability – **Art. 12(3)(b)&(c)**, **Art. 19**
Relevancy	Data is appropriate and applicable to the intended task.	Personal data shall be adequate, relevant, and limited to what is necessary in relation to the purposes – **Art. 5(1)(c)**, **Recital 39**	Requires assessment of the dataset's suitability – **Art. 10(2)(e)** Datasets must be relevant to the intended purpose – **Art. 10(3)**
Timeliness	Data is up-to-date and available when needed.	Personal data must be kept up to date – **Art. 5(1)(d)**	
Completeness	All necessary components of data are present and adequately recorded.	Adequate and relevant personal data for intended purposes – **Art. 5(1)(c)**	Datasets must be complete for the intended purpose – **Art. 10(3)** Completeness to avoid biased or inaccurate AI outputs – **Recital 67**
Documentation	Clear and comprehensive information should describe the data and its management.	Accountability principle mandates documented processing activities – **Art. 5(2)**, **Art. 35**, **Recital 82** Requires maintaining records of processing activities – **Art. 30(1), (3)&(4)**	High-risk AI systems shall be accompanied by instructions for the input data – **Art. 13(3)(b)(vi)** Record-keeping of input data that led to a match – **Art. 12** The technical documentation shall contain information on the data – **Art. 11**, **Art. 18**, **Annex IV**, **Annex XI**
Accessibility	Data can be accessed and used by authorised users.	Right of access for data subjects – **Art. 15** Data portability in structured formats (facilitating data accessibility) – **Art. 20**	
Security	Personal data should be protected from unauthorised disclosure or misuse.	Mandates integrity and confidentiality through robust safeguards – **Art. 5(1)(f)** Privacy by design/default embedded in systems development – **Art. 25** Requires appropriate technical and organisational measures to ensure a level of security appropriate to the risk – **Art. 32**	Privacy protection throughout AI lifecycle, adhering to the GDPR principles – **Recital 69** Incorporates privacy and data protection into its risk management and ethical frameworks for AI systems – **Recital 27** AI Act does not affect personal data privacy in the GDPR – **Art. 2(7)**
Bias mitigation	Data is managed to identify, reduce, and monitor biases, supporting more equitable treatment across groups.	Automated individual decision-making should not be based on special categories of personal data – **Art. 22(4)**, **Recital 71**	Data governance and management practices to detect, prevent and mitigate possible biases – **Art. 10(2)(f)&(g)** Datasets shall have appropriate statistical properties regarding persons or groups of persons – **Art. 10(3)** Exceptionally process special categories of personal data for bias detection and correction – **Art. 10(5)**, **Recital 70** Attention to bias mitigation in the data sets – **Recital 67**

"They ask you to show that your training data meets quality and compliance requirements. They want evidence of accuracy checks, bias assessments, and documentation of how the data were handled."

"What could your team realistically provide today?"

"What information would be difficult or impossible to show without additional work?"

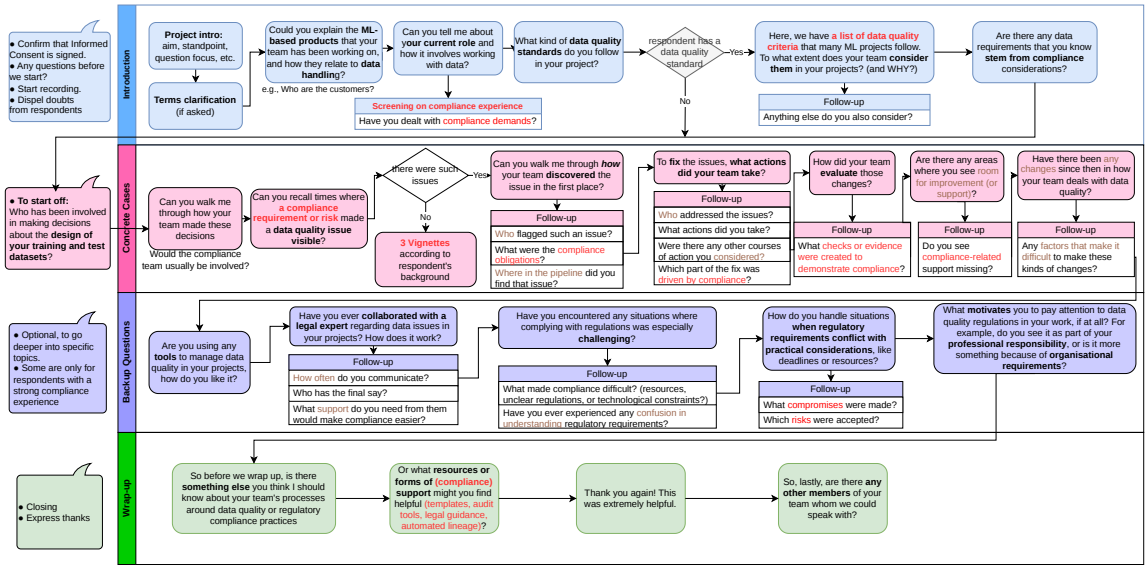

Fig. 2. Interview outline used in the semi-structured interviews. The outline guided discussions on data quality practices, compliance-related challenges, and regulatory-aligned trade-offs.

D Coding Structure and Analytical Process

D.1 Deductive Codebook

Table 4 presents the full deductive codebook.

D.2 Inductive Themes

Alongside deductive codes, inductive coding captured themes not anticipated by the initial framework. Table 5 presents inductively derived themes, their descriptions, and illustrative evidence drawn directly from participant accounts.

D.3 Interdisciplinary Refinement

Emerging themes were refined through iterative discussion among researchers with backgrounds in machine learning, data engineering, and law. This process served three purposes.

First, it ensured consistent interpretation across technical and legal framings. For example, the term *accuracy* means statistical correctness to engineers but carries a rectification obligation under GDPR Art. 5(1)(d). Without discussion, such terms risk being coded inconsistently.

Second, it surfaced cases where a single participant account implicated multiple codes. Deletion requests, for instance, simultaneously raised concerns about traceability, security, and data subject rights. These overlaps were noted explicitly rather than resolved by arbitrary single-code assignment.

Third, it resolved ambiguities between codes with similar surface forms. The compliance-engineering translation gap describes a knowledge and tooling problem: practitioners lack the means to map legal obligations to pipeline steps. Role boundaries describes a governance problem: ownership over compliance tasks is unclear between

teams. These are distinct, though they co-occur in practice. Similarly, compliance champion overload concerns resourcing, while legal involvement timing concerns process sequencing.

Disagreements were resolved through discussion until consensus was reached. No formal inter-rater reliability metric was computed, consistent with the flexible coding approach of Deterding and Waters [19], which treats coding as an interpretive process rather than a classification task.

Table 4. Deductive codes were derived from three sources: (1) established data quality dimensions, (2) relevant provisions in the GDPR and AI Act, and (3) the three research sub-questions. The systematic mapping is developed in a companion survey paper [80].

Code category	Code	Description
Data quality dimensions (SRQ1)	Accuracy	Correctness and reliability of data records. Engineering pursues quantitative precision; regulations require context-sensitive sufficiency and mandate rectification or erasure of inaccurate personal data.
	Completeness	Presence of all necessary data components for the intended purpose, extended by the AI Act to include population representativeness. Tension arises with GDPR data minimisation.
	Traceability	Tracking data origins, changes, and pathways to support accountability. Legally requires processing logs and automated event logging; in practice often reconstructed manually only at audit time.
	Relevancy	Appropriateness and necessity of data for the intended task. Engineering uses predictive value; the GDPR applies stricter necessity and proportionality standards that technically justified feature choices may not satisfy.
	Timeliness	Currency and availability of data when needed. Engineering tolerates some staleness; legal standards are context-driven and risk-based, sometimes requiring stricter update protocols.
	Documentation	Records describing data, processing decisions, and management practices. Technically supports reproducibility; legally serves as formal compliance evidence. In practice, records are often absent, outdated, or fragmented across tools.
	Accessibility	Ability of authorised users to access and use data. The GDPR extends this to data subject rights of access and portability; in practice, permission management creates both over-access and bottlenecks.
	Security	Protection of personal data through technical and organisational safeguards. Requires integrity, confidentiality, and privacy by design; in practice deprioritised under delivery pressure.
	Bias mitigation	Impartiality of data and equitable treatment across groups. Legal frameworks require proactive bias detection and mitigation beyond engineering fairness-accuracy trade-offs; subgroup gaps are often detected only post-deployment.
Tooling and infrastructure (SRQ2)	Existing tool use	Tools currently used for data quality management, including commercial platforms, open-source libraries, and custom solutions.
	Custom tooling	Internally built solutions compensating for off-the-shelf gaps; resource-intensive and largely limited to larger organisations.
	Tool limitations	Gaps between tool capabilities and regulatory needs; most tools target performance rather than legal compliance.
	Fairness tooling	Methods and tools for bias detection and subgroup evaluation; typically ad hoc scripts rather than standardised pipelines.
	Pipeline fragmentation	Data flows across tool and organisational boundaries with no end-to-end visibility, making consistent compliance structurally difficult.
Collaboration and governance (SRQ3)	Compliance–engineering translation gap	Legal obligations cannot be operationalised into concrete pipeline steps without additional interpretive work.
	Role boundaries	Clarity of ownership over compliance tasks between data and legal teams; in practice each party often assumes the other is responsible.
	Legal involvement timing	When compliance teams are consulted relative to pipeline development; late involvement forces costly rework under delivery pressure.
	Shared language	Consistency of meaning for terms such as accuracy, relevance, and bias across legal and technical teams; divergence produces coordination failures.
	Trade-off navigation	How practitioners balance engineering performance and compliance goals; compliance frequently loses to delivery timelines and management pressure.
	Motivation	Drivers of compliance practice: external pressure (audits, fines) versus internal values (responsibility, culture); produces different behaviours and varies by industry and organisational size.

Table 5. Inductively derived themes with descriptions and illustrative evidence.

Inductive theme	Description	Illustrative evidence
Deletion propagation failure	Personal data traces persist in derived features, historical training sets, and intermediate extracts even after removal from primary storage.	"If a user asks for deletion, how do we remove them from derived features or historical model training data? Retraining becomes expensive or infeasible." (R5)
Compliance debt accumulation	Compliance work is knowingly deferred under delivery pressure, creating latent regulatory risk that compounds as systems evolve.	R4 admitted running experiments with a legally uncertain dataset "just to see if it helps," aware it did not meet the GDPR standards for lawful basis.
Audit-triggered reconstruction	Lineage diagrams, documentation, and compliance evidence are created reactively only when regulators or auditors request them.	R3 noted their team only created lineage diagrams "when something goes wrong," and even then the process was manual and incomplete.
Compliance as one-time hurdle	Compliance is treated as a gate at initial design approval rather than a continuous practice, with teams assuming freedom to evolve systems between formal reviews.	Once legal approved the initial design, teams assumed they were "free to do whatever we want" until the next major review. (R11)
Distributed ownership vacuum	No single team or individual holds responsibility for the full data lifecycle, resulting in pockets of good practice but no end-to-end governance.	"The data starts in the upstream team, and by the time it reaches us, no one really knows who validated what." (R9)
Management pressure on provenance	Delivery timelines create organisational pressure to adopt third-party datasets before their lawful basis or collection conditions are verified.	Several participants admitted this pressure sometimes wins; R4 described running experiments with a dataset of uncertain provenance to "see if it helps."
Unclear fairness acceptance criteria	Fairness fixes lack agreed thresholds, forcing teams to make underdefined trade-offs between subgroup performance and overall model accuracy.	Fairness fixes rarely had clear acceptance criteria. This forced teams to weigh performance losses against the need to meet AI Act requirements. (multiple participants)
Post-deployment subgroup discovery	Subgroup performance gaps are detected only after deployment because evaluation pipelines focus on aggregate metrics during development.	Models "perform well on aggregate metrics but hide significant drops for specific groups", detected post-deployment. (multiple participants)
Security shortcuts under pressure	Broad data access is granted to speed up development, bypassing internal access control policy under delivery pressure.	Under delivery pressure, teams sometimes granted broad access even when this conflicted with internal policy. (multiple participants)
GenAI documentation limits	Generative AI tools can summarise code or metadata but cannot keep documentation synchronised with evolving pipelines, forcing teams back to manual editing.	R5 and R6 noted AI assistants could summarise but could not maintain live synchronisation with pipelines.
Compliance champion overload	Individuals who informally bridge legal and technical teams carry disproportionate workloads and lack formal authority or dedicated time, making the role unsustainable.	Internal compliance champions "carried significant workloads and spent considerable time translating regulatory guidance into actionable engineering steps."
Industry-based regulatory awareness variation	Familiarity with regulatory requirements varies systematically across industries, with finance and banking practitioners showing stronger baseline awareness than those in less regulated sectors.	Participants in finance and banking consistently reported higher familiarity with regulatory requirements, whereas those in less regulated sectors described compliance as more ad hoc (cross-case synthesis).
Small organisation compliance burden	Smaller organisations lack the resources to build custom compliance tooling or dedicated compliance roles, making regulatory alignment disproportionately harder.	"Only large organisations have the resources to do this." (R3)