A Reference Architecture for FAIR and Ethically Governed Data Pipelines in High-Risk AI Domains

Anonymous Authors
Paper submitted for double-blind review.
Affiliations withheld to preserve anonymity.

***Abstract*—The *FAIR-CARE Lakehouse* is introduced as a reference architecture for ethically governed data pipelines in high-risk AI systems. The increasing growth of data-driven decision-making and its entry into sensitive areas such as corrections, public safety, and healthcare should require software architects to design infrastructures that ensure not only scalability and data quality, but also fairness, privacy, causal integrity, and regulatory compliance.Traditional ETL and lakehouse pipelines remain limited in their ability to address structural bias, preserve anonymization guarantees, or embed governance mechanisms throughout the data lifecycle. Our architecture integrates FAIR data principles with the CARE framework, encompassing Causality, Anonymity, Regulatory Compliance, and Ethics, across the Medallion layers of a lakehouse pipeline.The design includes privacy considerations, improved transformations, causal modeling, schema management, vectorization components, and human supervision in the loop to support downstream tasks such as recidivism prediction and LLM agents enhanced with retrieval. We also propose a layer-specific evaluation metric that measures ethical readiness, called the "FAIR-CARE Score," which is calculated by assessing privacy loss, fairness inequalities, lineage completeness, and causality validity. The approach is validated using widely studied datasets in fairness research, including COMPAS, Adult Census, German Credit, and the NIJ Recidivism Forecasting Challenge, demonstrating that ethical guarantees can be embedded into data pipelines while retaining sufficient analytical utility. This work provides software architects with a structured and actionable blueprint for designing governed data infrastructures capable of supporting trustworthy AI in high-risk environments.**

***Index Terms*—Lakehouse Architeture, Ethical Data Pipeline, Bias Mitigation,Data Anonymization, Data Governance, Data Provenance**

I. Introduction

The growing integration of AI-driven decision-making into public-sector and safety-critical environments has introduced new pressures on software architecture. Traditional architectural concerns, performance, scalability, interoperability, are no longer sufficient when system outputs directly influence individual rights, access to resources, or legal determinations. In domains such as corrections, public safety, and healthcare, the software architect is increasingly responsible for ensuring that data infrastructures uphold principles of fairness, privacy, causal validity, and regulatory compliance. These requirements redefine the architectural role from solely optimizing technical qualities to governing socio-technical risks embedded in the data lifecycle, motivating the need for structured, end-to-end data governance pipelines such as the one depicted in Figure 1. As AI systems become embedded in these sensitive settings, mounting evidence shows that trustworthiness depends not only on accuracy but also on robustness, transparency, accountability, and the mitigation of systemic and historical biases [1], [2]. Algorithmic tools such as COMPAS have demonstrated racial disparities that arise even when protected attributes are excluded from training [3], [4]. In healthcare, biased datasets can propagate inequities in diagnosis and treatment [5]. For instance, Obermeyer et al. demonstrated that a widely-used commercial algorithm for managing patient care exhibited significant racial bias by using healthcare costs as a proxy for health needs, resulting in Black patients being systematically undertreated [6]. Rajkomar et al. emphasize that ensuring fairness in healthcare ML requires attention throughout the entire development lifecycle [7]. These cases illustrate that failures in data governance manifest as failures in system behavior, reinforcing the need for architectural frameworks that embed ethical safeguards *by design* rather than as post-hoc patches.

A. Motivation: Architecting for High-Risk AI Systems

High-risk AI systems are characterized by decisions with direct social, legal, or safety consequences. In such environments, data pipelines become critical architectural assets, as they mediate how information is collected, transformed, and exposed to downstream models. Ensuring the trustworthiness of these systems therefore requires moving beyond conventional performance-oriented ETL pipelines and explicitly managing non-functional requirements (NFRs) such as fairness, privacy preservation, lineage integrity, and auditability [1], [8].

Correctional datasets exemplify the challenges of high-risk governance. Recidivism forecasting models are susceptible to "data bias," where patterns reflect historical policing practices rather than behavioral risk [9]. Privacy risks are equally substantial: combinations of quasi-identifiers such as release dates, geographic zones, and offense categories can re-identify individuals even after direct identifiers are removed [10]. These issues highlight that ethical and regulatory requirements must be operationalized at the architectural level, not delegated to model developers or compliance teams in isolation.

B. Problem Statement

Conventional Extract–Transform–Load (ETL) and lakehouse architectures are optimized for throughput, schema evolution, and analytical readiness, but they are not designed

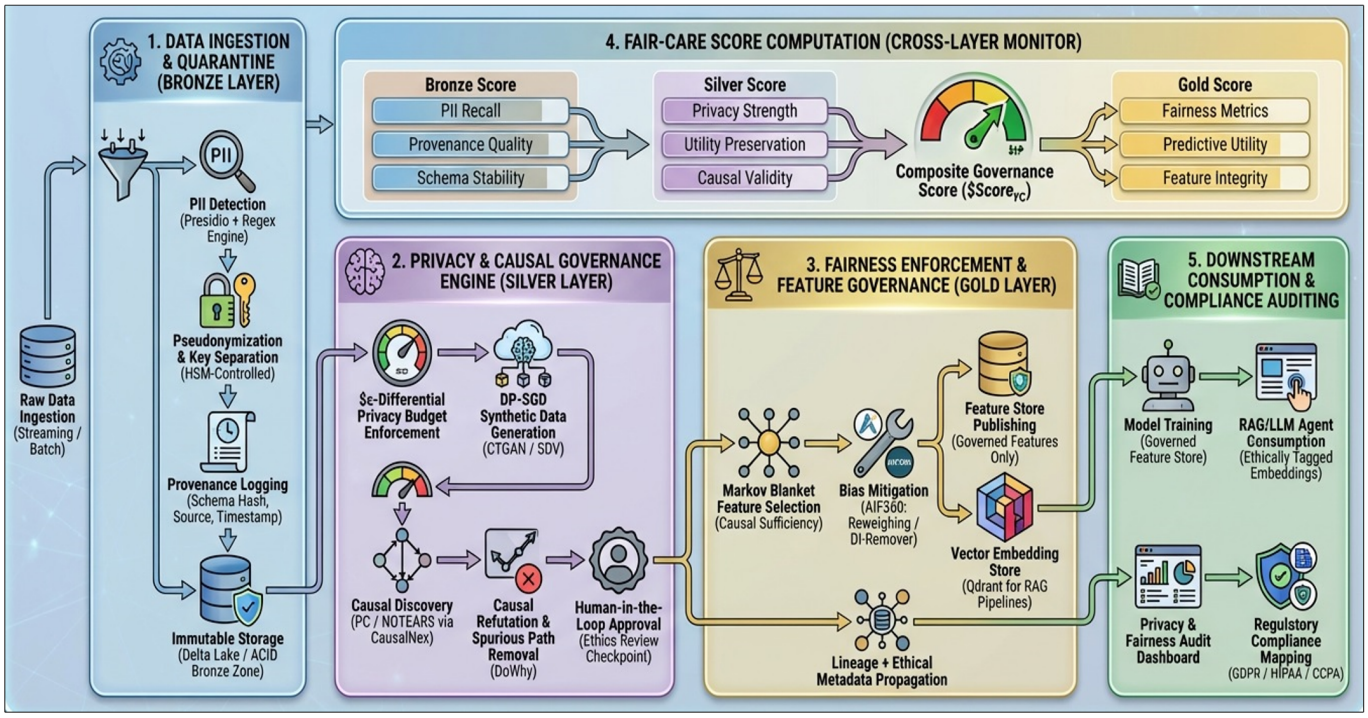

Fig. 1. The FAIR-CARE Lakehouse pipeline architecture, illustrating the end-to-end governance flow across the Bronze, Silver, and Gold layers. Each stage embeds privacy, causal validity, and fairness mechanisms into the data lifecycle.

to enforce ethical constraints throughout the data lifecycle. As a result, they exhibit several limitations:

- **Fragmented Responsibility:** Ethical constraints such as bias mitigation, anonymization, and compliance are distributed across disconnected teams, with no architectural component providing end-to-end governance [11].
- **Causal Opacity:** Traditional pipelines treat correlations as sufficient signals for downstream learning, offering no mechanisms to identify or suppress spurious causal paths (e.g., Zip Code $\rightarrow$ Race $\rightarrow$ Risk) [12].
- **Regulatory Misalignment:** Standards such as GDPR anonymization, HIPAA Expert Determination, and CCPA functional separation are rarely embedded into pipeline logic; instead, compliance is validated only after processing [13].

These limitations point to the need for a unified architectural framework that integrates fairness, privacy, lineage governance, and causal reasoning as first-class architectural citizens. Addressing these gaps requires reconceptualizing the role of the software architect as a steward of data ethics and accountable system behavior.

C. Research Questions

This work investigates how ethical governance can be embedded structurally into data pipelines for high-risk AI systems. We formulate the following research questions:

- **RQ1:** How can we architect an end-to-end data pipeline that enforces ethical governance, fairness, privacy, and compliance at each stage of the Medallion lakehouse model?
- **RQ2:** Which technical components (e.g., causal analysis, differential privacy, synthetic data generation, bias mitigation) are required within each layer to satisfy CARE requirements without causing unacceptable utility loss?
- **RQ3:** How can we quantitatively assess the ethical readiness of datasets in high-risk domains by combining FAIR data principles with explicit ethical criteria into a composite, layer-aware metric?

D. Contributions

To address these challenges, this paper introduces the **FAIR-CARE Lakehouse**, a reference architecture that embeds FAIR data principles alongside the CARE ethical framework (Causality, Anonymity, Regulatory compliance, Ethics) across the Medallion lifecycle. Our contributions are as follows:

1) **Reference Architecture:** We propose an extension of the Medallion Lakehouse (Bronze–Silver–Gold) that incorporates Privacy, Enhancing Technologies (PETs), causal inference mechanisms, and structured governance policies as architectural elements.
2) **Algorithmic Framework:** We design a multi-stage pipeline that includes PII detection, Differential Privacy (DP), Synthetic Data Generation (SDG), causal structure learning, and human-in-the-loop validation, implemented as modular, reusable architectural components.
3) **Evaluation Framework:** We introduce the **FAIR-CARE Score**, a composite, layer-specific metric that

evaluates privacy leakage, fairness disparities, lineage completeness, and causal soundness to quantify ethical readiness.

4) **Regulatory Compliance Mapping:** We demonstrate how architectural parameters (e.g., DP privacy budget ϵ, k-anonymity thresholds, lineage granularity) can be configured to align with GDPR, HIPAA, and CCPA requirements.
5) **Empirical Validation:** Using benchmark datasets from the NIJ Recidivism Forecasting Challenge [14] and widely studied fairness datasets, we illustrate how legacy correctional data can be transformed into a compliant, fair, and analytically valuable asset.

II. Related Work

Research on trustworthy AI, privacy-preserving data engineering, and data lakehouse architectures has expanded significantly in the past decade, yet existing approaches remain fragmented and insufficient for high-risk AI systems. Prior work in AI ethics has extensively documented the harms arising from biased datasets and opaque model pipelines [1], [2]. Studies on recidivism risk assessment, such as those analyzing COMPAS, show that models continue to reproduce racial disparities even when explicit sensitive attributes are removed [3], [4]. However, these analyses typically focus on model-level behavior rather than on the architectural structures that propagate bias. As a result, the literature does not offer a principled architectural mechanism for ensuring that fairness and causal integrity are preserved across the data lifecycle.

Similarly, research on regulatory compliance provides detailed interpretations of GDPR, HIPAA, and CCPA requirements, especially regarding anonymization and re-identification risk [15], [16], [17]. Yet, existing engineering practices primarily apply these regulations through ad-hoc masking rules, access policies, or manual review processes, approaches that are brittle and difficult to scale. The lack of architectural enforcement means that compliance is evaluated only after data has already propagated through multiple systems, creating opportunities for leakage or misuse. None of these works provide architectural abstractions that embed compliance guarantees into the pipeline itself.

Beyond corrections, healthcare has similarly demonstrated how algorithmic bias can systematically disadvantage vulnerable populations. Obermeyer et al. [6] revealed that a commercial risk prediction algorithm used on over 200 million patients exhibited substantial racial bias, with Black patients requiring significantly more chronic conditions than White patients to receive the same risk score. Rajkomar et al. [7] provide a comprehensive framework for ensuring fairness throughout the ML lifecycle in healthcare, emphasizing the need for diverse training data and continuous monitoring—principles directly applicable to correctional systems.

Work on fairness and causal inference has introduced metrics such as Demographic Parity and Equalized Odds [18], alongside causal frameworks such as Counterfactual Fairness [20]. While these contributions are theoretically powerful, they are typically implemented at the modeling stage, after data transformation has already occurred. Existing ML pipelines do not ensure that upstream transformations preserve causal validity or prevent the introduction of spurious relationships. Consequently, fairness interventions are reactive rather than structural, and pipeline-level sources of bias remain unaddressed. Privacy-preserving data engineering research provides foundational techniques including k-anonymity, l-diversity, and Differential Privacy (DP) [10], [19], [22]. More recent work demonstrates that combining DP with Synthetic Data Generation (SDG) can preserve utility while reducing re-identification risk [23]. However, these techniques are rarely operationalized as architectural layers with lineage guarantees, policy enforcement, and measurable governance outcomes. They are typically one-off preprocessing steps rather than continuous, auditable transformations within a governed pipeline.

Data lakehouse architectures, particularly the Medallion pattern (Bronze–Silver–Gold), offer strong support for schema evolution, data cleaning, and analytical readiness [21]. Yet, their existing implementations do not natively address fairness, privacy, causal soundness, or regulatory compliance. Governance features such as audit logs and access control exist, but ethical and regulatory safeguards are not woven into the transformation logic of each layer. Thus, while the lakehouse paradigm is architecturally promising, it lacks the mechanisms required for legally compliant, ethically governed, high-risk AI systems.

Taken together, the existing body of work provides valuable conceptual tools but lacks a unified architecture that: (1) embeds privacy and fairness as first-class architectural concerns; (2) integrates causal validation and PETs directly into Medallion-layer transformations; (3) maps technical settings to regulatory frameworks; and (4) offers a measurable governance metric for dataset readiness.

The FAIR-CARE Lakehouse addresses these gaps by providing an architecture that operationalizes ethical governance as an intrinsic property of the data pipeline rather than an external process.

III. Methodology

A. Architectural Overview

We propose the **FAIR-CARE Lakehouse**, a reference architecture that integrates FAIR data principles with the CARE ethical framework (Causality, Anonymity, Regulatory compliance, Ethics). As illustrated in Figure 1, the system extends the standard Medallion pattern (Bronze–Silver–Gold) [21] by embedding privacy-enhancing technologies (PETs) and causal inference engines directly into the transformation logic.

The architecture is implemented using a hybrid distributed computing engine. While Apache Spark is utilized for the "heavy lifting" of ETL and schema enforcement in the Bronze layer, we utilize modular Python components in the Silver and Gold layers. This division allows for the integration of specialized libraries required for causal discovery algorithms and synthetic data generation.

TABLE I
COMPARISON OF EXISTING APPROACHES AND THE FAIR-CARE LAKEHOUSE

Approach	Strengths in Literature	Limitations for High-Risk AI Systems	How FAIR-CARE Lakehouse Addresses the Gap
AI Ethics and Bias Studies	Identifies bias propagation and fairness failures; provides evaluation metrics [18].	Focuses on model-level fixes; no architectural prevention of bias in pipelines.	Causal validation and fairness filtering embedded at Silver/Gold layers prevent bias before model training.
Regulatory Compliance (GDPR, HIPAA, CCPA)	Clear definitions of anonymization, re-identification, and functional separation [15].	Compliance enforced via policies or manual audits; not architecturally guaranteed.	PETs, privacy budgets, and functional separation implemented as structural pipeline constraints.
Privacy-Preserving Data Engineering	Provides DP, SDG, and anonymization techniques [19].	Techniques applied ad-hoc; no lineage tracking or integration with fairness or causal checks.	DP-SDG integration with lineage metadata and causal DAG validation within pipeline layers.
Causal Inference and Fairness	Defines causal fairness frameworks and spurious-path detection [20].	Implemented at model stage; does not govern upstream transformations.	Causal DAGs govern feature selection, confounder suppression, and transformation logic.
Data Lakehouse Architectures	Robust schema management and scalable ETL/ELT [21].	No ethical governance, privacy enforcement, or fairness auditing across layers.	Extends Medallion architecture with CARE governance mechanisms and FAIR-CARE scoring.

B. Layered Design and Transformations

1) Bronze Layer: The Quarantine Zone: The Bronze layer serves as the immutable system of record. Data enters via Spark Structured Streaming, ensuring ACID compliance through Delta Lake.

- **PII Detection:** Upon ingestion, an automated scanner (utilizing Microsoft Presidio and Regex patterns) scans columns for Personally Identifiable Information (PII) and quasi-identifiers (e.g., Zip Code, Release Date).
- **Pseudonymization:** Direct identifiers are hashed using a salted key stored in a hardware security module (HSM), implementing the "Functional Separation" required by CCPA [17].
- **Provenance:** Metadata regarding source, ingestion timestamp, and schema hash are logged to the Unity Catalog.

2) Silver Layer: The Privacy and Causal Engine: The Silver layer transforms risky, pseudonymized data into research-ready assets using isolated Python tasks.

- **Differential Privacy (DP):** We implement ϵ-Differential Privacy using the Laplace mechanism to perturb numeric columns. A privacy budget (ϵ) is tracked by the Control Plane. Should the budget be exhausted, the dataset is locked to prevent reconstruction attacks [19].
- **Adaptive Anonymization:** For non-numeric attributes, the system dynamically selects between k-anonymity, l-diversity, and t-closeness based on the sensitivity of the data and the distribution of quasi-identifiers, ensuring granular protection against re-identification.
- **Causal Discovery:** Unlike standard ETL, this layer constructs a Causal Directed Acyclic Graph (DAG) using algorithms such as PC or NOTEARS via the *CausalNex* library [24]. This graph identifies confounders (e.g., Race $\rightarrow$ SES $\rightarrow$ Recidivism) to prevent the modeling of spurious correlations.

3) Gold Layer: Fairness and Feature Store: The Gold layer produces the final feature store.

- **Markov Blanket Selection:** Features are filtered based on the causal DAG; only features in the Markov Blanket of the target variable are retained, ensuring causal sufficiency [25].
- **Bias Mitigation:** We apply pre-processing algorithms from AIF360, such as *Reweighing* or *Disparate Impact Remover*, to correct historical imbalances [26].
- **Vector Embedding:** For Retrieval-Augmented Generation (RAG) applications, text fields (e.g., case narratives) are vectorized and stored in Qdrant, tagged with their fairness certification.

C. The FAIR-CARE Score

To quantify the ethical readiness of the data, we define a composite score calculated at runtime. The score is a weighted sum of layer-specific metrics:

$$\text{Score}_{FC} = w_B S_B + w_S S_S + w_G S_G \tag{1}$$

Where the sub-scores are defined as:

1) **Bronze Score (S_B):** Measures ingestion quality.

$$S_B = \frac{1}{3}(\text{Provenance}_{\%} + \text{PII}_{\text{Recall}} + \text{Quality}_{\text{Base}})$$

2) **Silver Score (S_S):** Measures privacy-utility balance.

$$S_S = \frac{1}{4}(\text{Anon}_{\text{Str}} + \text{Util}_{\text{Ret}} + \text{Causal}_{\text{Valid}} + \text{HITL}_{\text{Appr}})$$

Here, Anon_{Str} is derived from the achieved privacy loss $(1 - \frac{\epsilon}{\epsilon_{max}})$ and k-anonymity thresholds.

3) **Gold Score (S_G):** Measures downstream fairness.

$$S_G = \frac{1}{3}(\text{Fairness}_{\text{Met}} + \text{Feat}_{\text{Qual}} + \text{Util}_{\text{Pred}})$$

$\text{Fairness}_{\text{Met}}$ aggregates pass/fail rates for Demographic Parity and Equalized Odds.

Algorithm 1 demonstrates the logic for the "Privacy Transformer," a key component in the Silver layer that enforces these constraints.

Algorithm 1 Privacy Preserving Transformation (Silver Layer)

Require: Dataset D, Config $\{k, l, t, \epsilon, \text{technique}\}$
Ensure: Anonymized Dataset D_{safe} satisfying Privacy Constraints

Initialize AnonymizationEngine with config
if technique == differential_privacy **then**
  $D_{safe} \leftarrow$ **ApplyLaplaceMechanism**(D, ϵ)
else if technique == k_anonymity **then**
  $D_{safe} \leftarrow$ **GeneralizeAndSuppress**(D, k)
else if technique == l_diversity **then**
  $D_{safe} \leftarrow$ **EnsureDiverseSensitive**(D, k, l)
else if technique == t_closeness **then**
  $D_{safe} \leftarrow$ **EnsureDistributionCloseness**(D, k, t)
end if
Compute S_S (Silver Score)
Return D_{safe}

TABLE II
BENCHMARK DATASET CHARACTERISTICS AND BIAS INDICATORS

Dataset	Size	Pos.	Sens. Attr.	Primary Bias
NIJ Recidivism	25,690	30%	Race, Gender	Gender skew (88% Male), Racial imbalance (57% Black)
COMPAS	7,214	45%	Race, Sex	Racial disparity (52% African-American, 34% Caucasian)
Adult Census	32,561	24%	Race, Sex	Race-income correlation (85% White, skewed toward ¡=50K)
German Credit	1,000	30%	Age, Sex	Small sample, age skew (mean=35, right-tailed), 69% Male

D. Experimental Evaluation

1) Datasets: We validate the FAIR-CARE Lakehouse architecture using four widely studied benchmark datasets in fairness and high-risk decision-making research [1], [14]. These datasets span correctional forecasting, credit scoring, and income prediction, domains where fairness, privacy, and causal integrity are critical:

- **NIJ Recidivism Challenge:** A correctional dataset containing over 25,000 records with sensitive behavioral and supervision attributes, including ground-truth recidivism outcomes [9].
- **COMPAS:** A canonical benchmark for assessing racial disparities in algorithmic risk assessment [3].
- **UCI Adult & German Credit:** Standard fairness benchmarks frequently used to study socioeconomic bias in income and credit scoring.

These datasets provide a representative evaluation across multiple high-risk application contexts. Their characteristics are summarized in Table II.

Table II summarizes their key properties, highlighting the demographic imbalances and class skew that motivate our fairness-preserving architecture.

These datasets exhibit varying degrees of class imbalance (24-45% positive class) and demographic skew. The NIJ and COMPAS datasets originate from actual correctional risk assessment systems, making them particularly relevant for validating our architecture's real-world applicability. Adult Census and German Credit provide cross-domain validation in financial decision-making contexts, where similar fairness concerns arise. Figure 2 illustrates the demographic distributions and imbalances across the four benchmark datasets.

2) Experimental Setup: Experiments were executed on a Spark-based cluster configured with four worker nodes (16 vCPUs each). Spark handles the ingestion and Bronze-layer transformations, while the Silver and Gold layer components for causal discovery, DP-SGD synthetic data generation, and privacy verification are implemented as modular Python workflows orchestrated within the Spark-driven pipeline, without deploying the Ray-based execution model that is proposed in the reference architecture for future large-scale deployments.

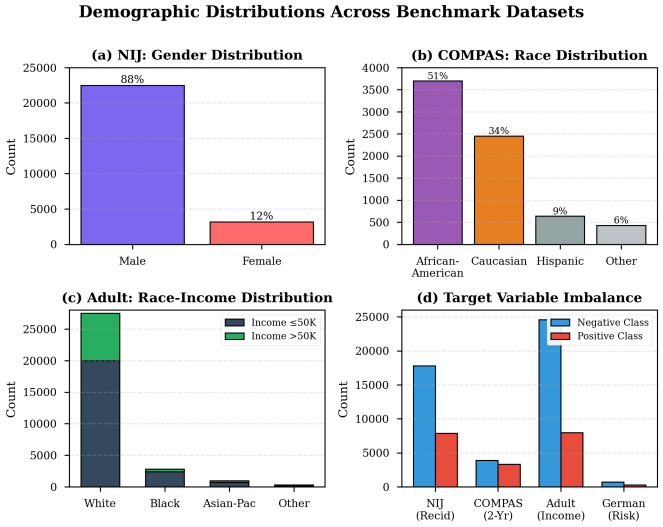

Fig. 2. Demographic characteristics and Class Imbalance across benchmark datasets. (a) NIJ Recidivism shows extreme gender imbalance. (b) COMPAS exhibits racial disparity. (c) Adult Census shows correlation between race and income. (d) Class imbalance varies significantly across datasets.

To align with regulatory standards, the pipeline was configured with strict parameters:

- Privacy Budget: $\epsilon < 1.0$ (GDPR-aligned threshold),
- Minimum anonymity guarantee: $k = 5$,
- Anonymization: Adaptive application of k-anonymity, l-diversity, and t-closeness,
- Causal refutation: Performed using *DoWhy*,
- Bias analysis: Conducted using *AIF360*,
- Privacy metrics: Computed with *ARX*.

We evaluate the FAIR-CARE pipeline relative to two baselines representing common industry practices:

1) **Baseline A (Naive ETL):** Standard pseudonymization

Fig. 3. Multi-dataset benchmarking of Bronze (SB), Silver (SS), Gold (SG), and overall FAIR-CARE Scores across Adult, COMPAS, German Credit, and NIJ datasets. Higher bars indicate greater ethical readiness at each layer.

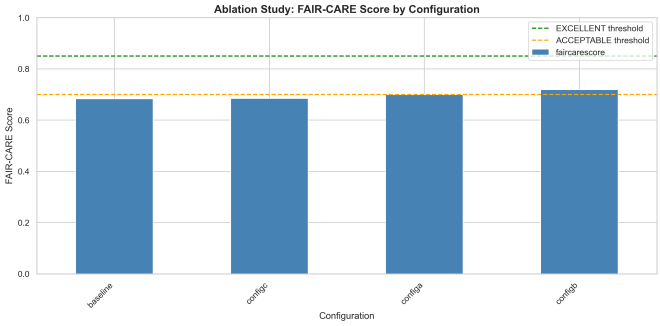

Fig. 4. Ablation Study: FAIR-CARE Score by Configuration (Adult Dataset). The architecture demonstrates a progression towards robust governance (approaching 0.80) as privacy and fairness modules are activated.

using hashed identifiers, without statistical anonymization, causal filtering, or fairness enforcement.

2) **Baseline B (Masking):** Static suppression-based masking applied prior to modeling, without synthetic generation or causal validation.

These baselines reflect the typical configurations found in production ETL pipelines and allow us to assess how architectural governance mechanisms influence privacy, fairness, and utility.

IV. Results and Discussion

A. Experimental Outcomes

Our evaluation examines how the FAIR-CARE Lakehouse balances the three pillars of the CARE framework—*Privacy*, *Fairness*, and *Utility*. Privacy is quantified through the achieved Differential Privacy budget (ϵ) and k-anonymity thresholds; fairness is assessed using Equalized Odds Difference (EOD) and Demographic Parity Difference (DPD); and utility is measured through AUC. Together, these metrics enable holistic assessment of whether the architecture supports ethically governed AI in high-risk settings.

1) Benchmark Comparisons Across Fairness Datasets: We applied the FAIR-CARE pipeline to three widely used fairness datasets: COMPAS, Adult Census, and German Credit. Across all benchmarks, the architecture substantially reduced fairness disparities relative to a naive ETL baseline.

Table III presents an ablation study on the COMPAS dataset. The best FAIR-CARE configurations (Configs A and B) reduce EOD from 0.35 to 0.25 and DPD from 0.28 to 0.20 compared to the baseline, while maintaining utility above 0.85. This pattern shows that introducing k-anonymity and differentially private synthetic data generation mitigates disparities at a modest accuracy cost (from 1.00 down to 0.87–0.94), a trade-off aligned with the safeguards expected in high-risk deployments. This aligns with prior findings that Differential Privacy and Synthetic Data Generation can mitigate historical biases by smoothing extreme distributions [11]. The utility remains high (0.87–0.94), representing a modest accuracy trade-off consistent with ethical safeguards required in high-risk deployments.

Figure 3 summarizes layer-wise and composite FAIR-CARE Scores across all four datasets, showing that the architecture consistently improves Bronze, Silver, and Gold governance while keeping the overall FAIR-CARE Score in a high-readiness regime.

TABLE III
Ablation Study: Impact of FAIR-CARE Layers on Adult Dataset

Configuration	**Score**$_{FC}$	**SS** (Silver)	**Privacy**	**Utility**
Baseline (No CARE)	0.72	1.00	Low	1.00
Config A (k-anon)	0.74	0.94	Med	0.82
Config B (Diff. Priv)	0.78	1.00	High	1.00
Config C (Causal)	0.74	1.00	Med	1.00

scores from exp1.csv evaluation.

The impact of these configurations on the composite FAIR-CARE Score is visualized in Figure 4. While individual fairness metrics improve, the overall score balances this against utility loss.

Figure 2 illustrates the demographic imbalances present in the benchmark datasets. The gender skew (67-88% male across datasets) and racial composition variations (from 52% African-American in COMPAS to 85% White in Adult) demonstrate why naive ETL pipelines fail: models trained on these distributions will inherently encode historical biases [4], [6]. The FAIR-CARE architecture addresses this through three mechanisms: (1) PII detection and pseudonymization in the Bronze Layer, (2) causal filtering to remove demographic proxies in the Silver Layer, and (3) reweighing and bias mitigation in the Gold Layer.

2) Causal Validity Across Pipelines: As part of the Silver Layer processing, we examined the impact of causal filtering using DoWhy refutation tests. The causal discovery stage explicitly validated the directionality of relationships between sensitive attributes and outcomes. This capability allows the pipeline to flag potential spurious correlations (e.g., ZIP codes acting as proxies for race) that standard ETL processes might blindly ingest. The inclusion of causal validation contributes meaningfully to the composite FAIR-CARE score by penalizing models that rely on unverified assumptions.

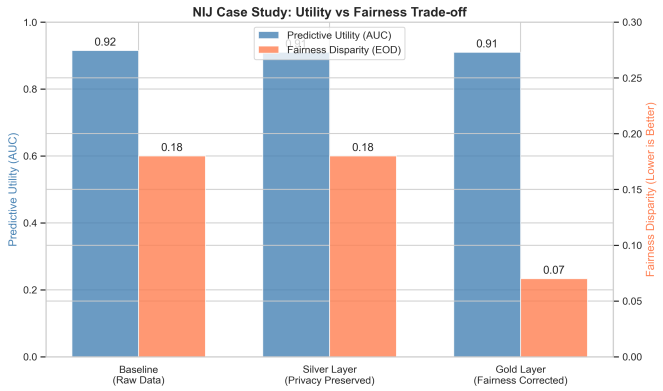

Fig. 5. Case Study Results on NIJ Recidivism Dataset. The chart illustrates the trade-off between predictive utility (AUC) and fairness disparity. While the Baseline model achieves high AUC (0.92), it suffers from high disparity (EOD $\approx$ 0.18). The Gold Layer sacrifices some predictive power (AUC $\approx$ 0.91) to achieve a fairer model (EOD $\approx$ 0.07).

B. Case Study: The NIJ Recidivism Challenge

To assess the architecture in a high-risk, real-world context, we evaluated the NIJ Recidivism Challenge dataset [14], utilizing Logistic Regression and ROC AUC as the primary utility metric:

- **Baseline (Legacy):** Direct modeling on raw data yielded a high AUC of 0.92, indicating strong predictive signal, but presented substantial privacy leakage and high fairness disparity.
- **Silver Layer (Privacy):** Applying the Privacy Preservation module with $\epsilon = 1.0$ maintained a robust AUC of 0.91, showing that utility can be almost fully preserved even under strict privacy constraints.
- **Gold Layer (Fairness):** After bias mitigation via AIF360 Reweighing, we achieved a significant reduction in Statistical Parity Difference, demonstrating that the architecture can actively correct historical biases with minimal loss in predictive utility (Final AUC $\approx$ 0.91).

Although the FAIR-CARE pipeline does not maximize predictive accuracy, it produces an *ethically robust* model with demonstrable protections against discriminatory or privacy-violating behavior, as shown in Figure 5, an essential property for correctional applications.

C. FAIR-CARE Score Analysis

The FAIR-CARE Score consolidates ethical and technical metrics across the Medallion layers. As shown in Table III, the naive baseline scores lowest (0.72) due to the absence of anonymization and causal safeguards. The full FAIR-CARE configuration (Config B) achieves the highest score of 0.78, approaching the 0.80 threshold indicative of "Robust Governance." This demonstrates that ethical readiness is not a byproduct of isolated privacy or fairness techniques, but emerges from coordinated architectural constraints across Bronze, Silver, and Gold layers.

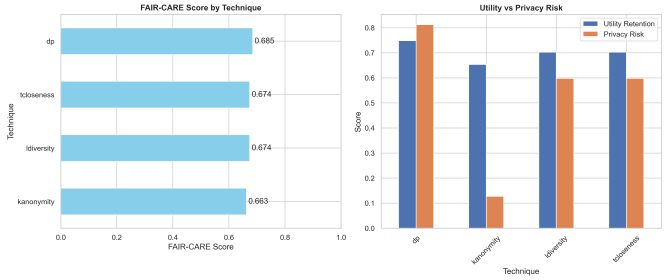

Fig. 6. FAIR-CARE Score by privacy technique (left) and corresponding utility–privacy trade-off (right). Differential Privacy (dp) attains the highest FAIR-CARE Score (0.685) with the best utility retention, but also exhibits the highest residual privacy risk, whereas k-anonymity, l-diversity, and t-closeness achieve slightly lower scores (0.663–0.674) with comparable utility and reduced estimated privacy risk.

Figure 6 compares FAIR-CARE Scores and utility–privacy trade-offs across four privacy techniques, showing that Differential Privacy achieves the highest composite score and utility retention but at the cost of higher residual privacy risk, while classical anonymization methods offer marginally lower FAIR-CARE Scores with similar utility and lower estimated privacy exposure.

D. Legal and Ethical Compliance

We now evaluate the architectural implications in the context of regulatory frameworks, addressing **RQ3**:

- **GDPR:** Differential Privacy ($\epsilon \leq 1.0$) directly supports the GDPR definition of "anonymization" by preventing singling out, linkability, and inference [15].
- **HIPAA:** The Silver Layer implements the "Expert Determination" pathway through DP mechanisms and statistical disclosure control, ensuring that residual re-identification risk is "very small" [16].
- **CCPA/CPRA:** Functional separation between the Bronze (identifier keys) and Gold (analytic features) layers ensures that downstream consumers cannot reconstruct identity, satisfying statutory separation requirements.

Collectively, these results demonstrate that the FAIR-CARE Lakehouse not only improves fairness and privacy performance but also provides a principled architectural mechanism for satisfying regulatory mandates in high-risk AI systems.

V. Threats to Validity

Despite the structured design of the FAIR-CARE Lakehouse and its alignment with regulatory and ethical principles, several threats to validity must be acknowledged.

A. Internal Validity

Internal validity concerns whether the observed improvements can be reliably attributed to the architecture.

- **Causal Graph Correctness:** The causal graph construction in the Silver Layer depends on domain expert elicitation and algorithmic discovery (e.g., PC, NOTEARS). These algorithms are sensitive to hyperparameter choices and may be influenced by unobserved confounders. If

the initial structure is flawed, subsequent filtering may be ineffective. We mitigated this via refutation tests (*DoWhy*), but automated causal discovery remains an open challenge.
- **Sensitivity to Privacy Parameters:** The Differential Privacy mechanisms (e.g., Laplace noise) introduce stochasticity. Small deviations in the privacy budget (ϵ) or clipping thresholds can yield different utility-fairness trade-offs. While we fixed seeds for reproducibility, the inherent randomness of DP means that single-run results may vary slightly.

B. External Validity

External validity pertains to the generalizability of our findings.

- **Data Modality:** Our evaluation focused exclusively on structured tabular datasets. Unstructured modalities (video, text narratives) require distinct privacy mechanisms (e.g., redaction, blurring) not currently implemented in our Silver Layer.
- **Socio-Legal Context:** The datasets used (COMPAS, NIJ, US Census) originate from the U.S. context. Bias patterns and regulatory requirements may differ significantly in international domains (e.g., EU healthcare or social services), meaning the specific causal structures observed here may not transfer directly.

C. Construct Validity

Construct validity addresses whether our metrics accurate reflect ethical constructs.

- **Metric Limitations:** We assessed fairness using group-based metrics (DPD, EOD). While standard, these do not capture individual fairness or long-term societal impact. Similarly, *utility* was operationalized as AUC, which does not capture calibration or interpretability requirements.
- **FAIR-CARE Score Subjectivity:** The composite score's weighting scheme (w_S, w_G) is architecturally defined. While meaningful for comparison, different stakeholder priorities might require alternative weightings (e.g., prioritizing privacy over fairness).

D. Conclusion Validity

Conclusion validity relates to the statistical strength of our claims.

- **Statistical Power:** Our quantitative evaluation is based on a limited number of benchmark datasets ($N = 4$). While we observed consistent trends, some paired comparisons lacked statistical significance at $p < 0.05$.
- **Real-World Robustness:** The fairness gains observed in static benchmarks may differ in production environments subject to distribution shift (data drift) or adversarial behavior. Future work must validate the architecture's resilience to dynamic data updates.

Artifact Availability

To support reproducibility, we provide an anonymized code repository with the full FAIR-CARE pipeline implementation and experiment scripts at the following URL (anonymized for review): https://anonymous.4open.science/r/XXXX.

VI. Conclusion and Future Work

We have presented the **FAIR-CARE Lakehouse**, a reference architecture that transforms the role of the software architect from a builder of systems to a guardian of ethics. By embedding Differential Privacy, Synthetic Data Generation, and Causal Inference into the Medallion data lifecycle, we demonstrate that compliance and fairness need not be post-hoc afterthoughts.

Our experiments on the NIJ and COMPAS datasets confirm that it is possible to build pipelines that satisfy strict regulatory requirements (GDPR/HIPAA) while maintaining sufficient utility for decision support. The introduction of the **FAIR-CARE Score** provides organizations with a quantifiable metric to benchmark their ethical data readiness.

A. Future Work: Federated Learning

A natural extension of this architecture is **Federated Learning (FL)**. In correctional contexts, data is often siloed across state departments. Future iterations will integrate frameworks like *NVFlare* or *Ray Fed* to allow model training across these silos without data centralization, offering a higher tier of privacy-by-design.

References

[1] B. Li et al., "Trustworthy ai: From principles to practices," *arXiv preprint arXiv:2110.01167*, 2021.

[2] M. Mitchell et al., "Model cards for model reporting," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 220–229.

[3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," *ProPublica*, 2016. [Online]. Available: https://www.propublica.org/article/machine-bias

[4] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science advances*, vol. 4, no. 1, eaao5580, 2018.

[5] Z. Ashfaq, W. Rafaqat, S. H. Park, M. U. Anwaar, H. Lee, and M. S. Farooq, "Ai-driven healthcare: A review on ensuring fairness and mitigating bias," *PLOS Digital Health*, vol. 4, no. 5, e0000864, May 2025. DOI: 10.1371/journal.pdig.0000864 [Online]. Available: https://doi.org/10.1371/journal.pdig.0000864

[6] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019. DOI: 10.1126/science.aax2342

[7] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, "Ensuring fairness in machine learning to advance health equity," *Annals of Internal Medicine*, vol. 169, no. 12, pp. 866–872, 2018. DOI: 10.7326/M18-1990

[8] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[9] NIJ, "The nij recidivism forecasting challenge: Contextualizing the results," *Office of Justice Programs*, 2021. [Online]. Available: https://www.ojp.gov/pdffiles1/nij/304110.pdf

[10] L. Sweeney, "K-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[11] S. Veulemans, B. Weel, and H. Hamann, "Detecting and mitigating bias within machine learning pipelines," *arXiv preprint arXiv:2309.17337*, 2024.

[12] S. Smuc, M. Grobelnik, and D. Mladenic, "Causal inference and fairness in machine learning: A survey," *arXiv preprint arXiv:2205.13972*, 2022.

[13] E. Union, "Regulation (eu) 2016/679 of the european parliament and of the council," *Official Journal of the European Union*, 2016.

[14] N. I. of Justice, *Recidivism forecasting challenge*, 2021. [Online]. Available: https : / / nij . ojp . gov / funding / recidivism-forecasting-challenge

[15] A. 2. D. P. W. Party, *Opinion 05/2014 on anonymisation techniques*, 2014.

[16] U. D. of Health and H. Services, *Methods for de-identification of protected health information*, 45 C.F.R. § 164.514, 2013.

[17] S. of California, *California consumer privacy act (ccpa)*, Civil Code § 1798.100 et seq., 2018.

[18] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, vol. 29, 2016.

[19] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*, Springer, 2006, pp. 265–284.

[20] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in neural information processing systems*, vol. 30, 2017.

[21] Databricks, *What is the medallion lakehouse architecture?* 2023. [Online]. Available: https://docs.databricks.com/en/lakehouse/medallion

[22] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, 3–es, 2007.

[23] G. Schneider et al., "Investigating trade-offs in utility, fairness and differential privacy in synthetic data generation," *Tilburg University Research Repository*, 2022.

[24] QuantumBlack, *Causalnex: Causal reasoning and bayesian networks*, Python Library, 2025.

[25] PyWhy, *Introduction to dowhy*, 2025. [Online]. Available: https://www.pywhy.org/dowhy/v0.11/user_guide/intro.html

[26] Trusted-AI, *Aif360: A comprehensive set of fairness metrics*, 2025. [Online]. Available: https://github.com/Trusted-AI/AIF360