

TOPICAL REVIEW

Fairness and Explanations in Entity Resolution: An Overview

TIAGO BRASILEIRO ARAÚJO^{ID 1,2}, VASILIS EFTHYMIU^{ID 3}, AND KOSTAS STEFANIDIS^{ID 1}

¹Faculty of Information Technology and Communication Sciences, Tampere University, 33100 Tampere, Finland

²Federal Institute of Paraíba, Soledade 58051-900, Brazil

³Department of Informatics and Telematics, Harokopio University of Athens, 176 76 Athens, Greece

Corresponding author: Tiago Brasileiro Araújo (tiago.brasileiroaraujo@tuni.fi)

ABSTRACT Entity Resolution (ER) is a fundamental task in data integration, enabling the identification of records that refer to the same real-world entity across diverse and often heterogeneous data sources. Recent advances in Artificial Intelligence (AI) have significantly improved ER performance, particularly with deep learning and pre-trained embeddings. However, these AI-driven solutions introduce new challenges related to fairness and explainability. Fairness-aware ER seeks to mitigate bias that may arise from algorithmic decision-making or imbalanced training data, while eXplainable Entity Resolution (XER) aims to enhance transparency and trust in ER. In this work, we provide a comprehensive overview of fairness and explainability in ER, systematically analyzing existing techniques across the ER pipeline. We discuss challenges in ensuring unbiased and interpretable ER outcomes, with a special focus on streaming environments, where real-time decision-making intensifies the complexity of these concerns. Furthermore, we outline research opportunities and examine the trade-offs between schema-aware and schema-agnostic methods, as well as rule-based and machine learning-based comparison techniques, in ensuring fair and transparent ER. Our study highlights open research challenges and potential future directions, encouraging novel explainable AI methodologies and fairness-aware ER solutions that enhance the reliability, accountability, and societal impact of AI-driven ER systems.

INDEX TERMS Artificial intelligence, entity resolution, explainable artificial intelligence, fairness, streaming data.

I. INTRODUCTION

In the Big Data era, businesses, governments, and scientific organizations increasingly rely on vast amounts of data collected from different sources for supporting their decision making processes. However, those data sources often suffer from quality issues such as incompleteness (missing attributes), redundancy (overlapping records), inconsistency (conflicting values), or incorrectness (data errors) [1]. Entity Resolution (ER) is an important data integration task that seeks to identify different descriptions that refer to the same real-world entity, ensuring more reliable and unified datasets [2], [3]. Traditionally, ER has been applied to structured data, matching records within the same relational table

(deduplication) or across multiple tables (record linkage). However, in modern scenarios, ER also applies to semi-structured data, such as RDF knowledge bases, JSON files, and large-scale web data [4], [5].

The importance of ER extends to a wide range of real-world applications. In healthcare, ER is essential for linking patient records across hospitals, ensuring accurate patient history tracking. In airline security, it plays a crucial role in matching passenger records against no-fly lists, enhancing border security [6]. In e-commerce, ER enables retailers to detect fraudulent knockoffs and consolidate product information across different platforms. Similarly, social media platforms leverage ER to merge duplicate profiles and identify shared events across multiple sources [2].

Despite its long history and importance, ER remains highly challenging, particularly due to data heterogeneity,

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

schema variability, and scalability concerns. As data volumes continue to grow, traditional rule-based ER methods struggle with efficiency, leading to the adoption of Artificial Intelligence (AI) techniques, including deep learning and pre-trained embeddings, to improve accuracy and automation. However, the increasing reliance on closed-box AI models introduces new concerns regarding interpretability, fairness, and real-time adaptability, especially in streaming ER scenarios, highlighting the need for more explainable and ethical solutions in the field.

The rise of streaming data sources, such as social media platforms, IoT networks, and financial transaction systems, has amplified the necessity for ER techniques capable of processing data incrementally [7]. Traditional batch-based approaches struggle to meet the latency and scalability demands of these dynamic environments, leading to a surge of research on streaming ER methods.

Over the years, the evolution of ER techniques has paralleled the advancements in AI, particularly with the advent of deep learning models and pre-trained embeddings [8]. Therefore, the integration of AI in ER has significantly transformed traditional approaches, enhancing scalability and accuracy in matching scenarios. Machine learning models, particularly deep learning and graph-based techniques, have enabled ER systems to move beyond rule-based methods by learning entity representations and improving similarity computations [8], [9]. The use of pre-trained embeddings has further refined ER tasks, capturing contextual relationships across heterogeneous data sources. However, despite these advancements, AI-driven ER remains challenging, particularly in dynamic environments where entities continuously evolve. Ensuring efficiency in streaming ER requires models capable of incremental learning, adapting to new entity descriptions without retraining from scratch [10], [11]. Furthermore, as ER systems become more complex, the need for explanation and fairness-aware AI techniques has emerged as a critical research direction [12].

Alongside these technical challenges, ethical considerations, particularly related to fairness, have gained significant attention [13]. Biased entity matching can perpetuate societal inequalities, especially in sensitive applications like credit scoring, recruitment, and healthcare [6]. Consequently, fairness-aware ER approaches involve mitigating discrimination and bias, which can arise from the algorithm's design or from biased input data, such as misrepresented demographic groups [6], [13]. Ensuring fairness requires moving beyond accuracy optimization to incorporate constraints that address disparities between protected and non-protected groups. Key challenges include identifying protected attributes, like race or gender, and selecting appropriate fairness measures, such as equal representation or balanced error rates [14]. For example, in airline security, matching records from a no-fly list with passenger lists must consider demographic differences to prevent discriminatory outcomes. Despite its importance, fairness-aware ER remains an

underexplored area with open challenges for future research [6], [15].

Explainable Artificial Intelligence (XAI) has recently gained prominence as a means to enhance transparency, trust, and interpretability in AI-driven applications [16], [17]. In the context of ER, XAI techniques offer invaluable insights into the decision-making process by elucidating the factors that influence entity matches and mismatches. This lack of transparency not only undermines trust in automated decisions but also hinders the identification and mitigation of biases [18]. While prior works have addressed ER techniques from various angles, such as matching efficiency, deep learning architectures, or blocking strategies, none have focused on the joint integration of algorithmic fairness and explainable AI within ER pipelines. In this paper, we introduce the term eXplainable Entity Resolution (XER) to describe this integration of fairness aspects and XAI into ER processes. This integration offers the potential to enhance model interpretability, simplify error analysis, and support compliance with regulatory requirements.

Real-world applications increasingly highlight the importance of fairness and explainability in ER. Consider a national healthcare platform that integrates patient data from multiple hospitals, labs, and insurance providers. ER works in linking fragmented records referring to the same patient. However, without fairness-aware mechanisms, the system may disproportionately fail to match records of underrepresented populations, leading to biased health analytics or denial of benefits. Additionally, as these decisions are often unclear, healthcare administrators and affected individuals may not understand why certain records were not linked. An explainable ER pipeline could help expose which attributes led to a mismatch, enabling human auditors to assess and rectify the linkage. This scenario exemplifies the critical role of XER in real-world applications where decisions affect people in terms of access to services, rights, and opportunities.

This work delves into the intersection of ER, fairness, and XAI, highlighting the current state of the art, identifying open research challenges, and proposing potential directions for future work. Therefore, this paper contributes by (i) providing a comprehensive analysis of existing ER methods, with a focus on fairness-aware and explanation-driven approaches, (ii) addressing the concept of XER to bridge the gap between AI-driven ER and interpretable decision-making, (iii) examining open challenges in fairness-aware ER across all pipeline steps, and (iv) proposing future research directions, such as fairness-aware approaches and explanation models for ER.

The remainder of this paper is structured as follows. Section II describes the systematic literature review methodology adopted in this study. Section III introduces ER, outlining its key steps (i.e., blocking, matching, and clustering) along with their technical and computational challenges. Section IV provides foundational background on fairness and

explainability. Section V discusses fairness in ER, detailing how bias manifests in different ER stages and exploring fairness-aware approaches. Section VI focuses on XAI in ER, examining explanation techniques and their role in enhancing transparency and accountability. Finally, Section VII summarizes the findings, highlights open challenges, and proposes future research directions in fairness-aware and explainable ER.

II. METHODOLOGY

To provide a transparent and comprehensive overview of fairness and explainability in ER, this study adopts a Systematic Literature Review (SLR) methodology. Following structured review guidelines, the process includes defining research questions, selecting databases and keywords, and applying clear inclusion and exclusion criteria. This approach ensures balanced coverage of foundational and recent works across key ER dimensions, such as batch vs. streaming, schema-aware vs. schema-agnostic, and rule-based vs. ML-based methods. The methodology establishes a foundation for the comparative analyses and insights presented in the survey.

A. SEARCH STRATEGY

To ensure a comprehensive and focused review, we adopted a systematic search strategy inspired by established guidelines for conducting systematic literature reviews. Our goal was to identify relevant studies at the intersection of ER, Fairness, and XAI. The search was guided by a set of well-defined keywords, including: “Entity Resolution”, “Record Linkage”, “Entity Matching”, “Fairness”, “Bias Mitigation”, “Explainable Artificial Intelligence”, “Explanation”, “Interpretability”, “XAI”, “Transparency”, “Matching”, and “Blocking”. These terms were combined using Boolean operators to construct queries such as “Entity Resolution AND Fairness” and “Record Linkage OR Explainability”.

The literature search was carried out across multiple reputable scholarly databases, including ACM Digital Library, IEEE Xplore, SpringerLink, DBLP, and Google Scholar. These sources were selected due to their strong presence in publishing high-quality, peer-reviewed research in artificial intelligence and data management. We focused on publications from the period 2019 to 2025 to reflect the most recent advancements and trends in the field. However, we made exceptions for a few essential contributions that are widely cited and remain foundational to the current research landscape. To complement the initial search results, we also employed *backward snowballing*, reviewing the references of selected papers, as well as forward citation tracking, which allowed us to identify influential follow-up studies that may not have appeared directly through keyword-based search.

B. INCLUSION AND EXCLUSION CRITERIA

To ensure quality and relevance, this review applied clear inclusion criteria focused on studies addressing ER with fairness or explainability. Selected works include: *i*) novel methods or frameworks involving XAI or fairness;

ii) empirical evaluations highlighting explainability or bias mitigation; and *iii*) surveys or theoretical analyses on fairness and XAI in ER. Additionally, foundational ER studies (e.g., surveys, benchmarks, and overviews) were included to contextualize how fairness and explainability fit within the broader evolution of the field.

Exclusion criteria removed studies that: *i*) addressed generic ER without relevance to fairness or explainability; *ii*) discussed these themes broadly in AI with no ER focus; or *iii*) were non-peer-reviewed, duplicates, or lacked methodological depth. Works unrelated to ER were also excluded unless they clearly intersected with fairness or explainability in ER. The filtering process involved title and abstract screening, followed by full-text analysis to ensure alignment with the scope of this review.

C. STUDY SELECTION

The selection process aimed to ensure relevance and representativeness of studies on ER, fairness, and explainability. After compiling results from selected databases and removing duplicates, studies were screened by title and abstract, excluding those unrelated to ER or lacking content on fairness or explanation mechanisms. After a full-text review was conducted to confirm alignment with the inclusion criteria. A structured data extraction process was employed, collecting bibliographic details and classifying each study across analytical dimensions such as ER processing mode, comparison method, similarity evidence, fairness orientation, explanation approach, ER step addressed, use of pre-trained models, and scalability considerations.

These extracted elements guided the development of the taxonomies and comparative analyses presented throughout the article. Specifically, the selected studies are organized into tables and taxonomical frameworks in the subsequent sections, structured according to the criteria established in this methodology. This approach not only supports a comprehensive synthesis of the current literature but also supports the identification of open research challenges and opportunities for future work.

III. ENTITY RESOLUTION

ER, also known as entity matching, record linkage, or deduplication, is a fundamental process in data management aimed at identifying and matching records referring to the same real-world entity across one or more data sources [2]. Traditional ER workflows consist of blocking, matching, and clustering [1], as illustrated in Figure 1. Blocking groups similar records to reduce unnecessary comparisons while preserving recall. Matching measures the similarity between record pairs using predefined functions or machine learning models. Clustering groups together identified matches to ensure that all descriptions within a cluster correspond to the same entity in the real world. These steps collectively optimize ER by balancing efficiency and accuracy.

The ER process is inherently complex due to several factors. Data sources often lack a unified schema, resulting



FIGURE 1. Traditional ER workflow.

in heterogeneous representations of entities. For instance, names, addresses, or product descriptions may be formatted differently, contain errors, or exhibit missing values [19]. Additionally, the volume of data poses a significant computational challenge, as a naive approach to ER would require comparing every pair of records, leading to quadratic complexity. This issue is particularly critical in Big Data scenarios, where datasets can include millions of records [20].

The importance of ER extends beyond its immediate applications, serving as a foundational tool for various analytical and operational processes. Despite significant advances, challenges such as scalability, data quality, and ethical considerations continue to drive research in this domain [21], [22]. Overall, Figure 2 presents the taxonomy of ER settings based on their key characteristics.

A. BATCH VS. STREAMING PROCESSING

First, ER tasks can be executed in two primary processing modes: batch and streaming [2], [3], [23]. In batch processing, the system operates on a complete dataset that is fully available before the resolution process begins. This approach is effective for scenarios with static or infrequently updated data sources, where computational resources can be allocated for exhaustive processing. However, it becomes impractical in dynamic environments, where data arrives continuously. Streaming data, on the other hand, demands real-time processing to match entities as they are generated or updated. This dynamic context is typical of sources such as social media, IoT devices, and web platforms [24]. Incremental processing methods have been developed to manage this continuous flow of data while preserving a consistent view of previously matched entities. These scenarios underscore the need for scalable, efficient, and adaptive ER solutions capable of addressing the velocity, volume, and variety of streaming data [25].

Regarding streaming, works such as [25], [26], and [27] provided focus on streaming ER, addressing different aspects of efficiency and effectiveness in real-time record linkage. In the work [25], an incremental blocking technique is designed to handle heterogeneous and noisy data in streaming environments. The authors introduce two key strategies (attribute selection and top- n neighborhood entities) to optimize resource consumption while improving blocking efficiency. This method ensures that only the most relevant attributes contribute to blocking, reducing redundancy, and enabling efficient processing in parallel distributed environments.

In [26], the authors introduce ExpBlock, a randomized, memory-efficient blocking structure specifically designed

for streaming ER. Unlike traditional blocking techniques that rely on fixed rules or manually defined thresholds, ExpBlock employs probabilistic mechanisms to determine which records should be retained or discarded dynamically. This technique prioritizes frequently accessed and recently used blocks, exponentially decreasing the probability of retaining inactive records. By doing so, ExpBlock optimizes memory usage while ensuring accurate record linkage in high-velocity data streams. As an extension of [26], the authors of [27] present a set of randomized algorithms tailored for streaming ER. These algorithms leverage a bounded in-memory data structure, limiting both the number of blocks and the positions within each block. This ensures that the most frequently accessed and most recently used blocks remain in memory while records within each block are continuously updated. This approach is important for streaming data environments, ensuring that only the most relevant and frequently accessed data is retained while outdated or inactive records are seamlessly replaced.

Regarding explanation and fairness, the ER processing mode presents distinct challenges. In batch processing, models have access to the entire dataset, enabling the application of comprehensive explanation techniques, such as global feature attributions and counterfactual analysis. However, ensuring fairness can be complex, as biases embedded in historical data may propagate through the ER process. Addressing these biases requires fairness-aware pre-processing and post-hoc analysis to ensure equitable representation across demographic groups. In contrast, streaming ER demands real-time decision-making, where explanations must be incremental and adaptive. Techniques like local feature attribution and instance-based explanations become essential for providing timely insights. Fairness challenges are intensified, as data distributions can evolve over time, leading to fairness drift. Ensuring fairness in this context requires dynamic mechanisms that continuously monitor and adjust for bias. Furthermore, error propagation in streaming can accumulate, reinforcing unfair outcomes if early biases are not corrected promptly.

B. SCHEMA-AWARE VS. SCHEMA-AGNOSTIC TECHNIQUES

Another key dimension of ER is its adaptability to the similarity evidence, particularly with respect to schema dependency. Traditional schema-based ER relies on predefined structures and attribute mappings to match records, offering high accuracy in structured environments. However, it struggles with semi-structured or heterogeneous datasets where schemas vary or are absent, such as data sourced from diverse web

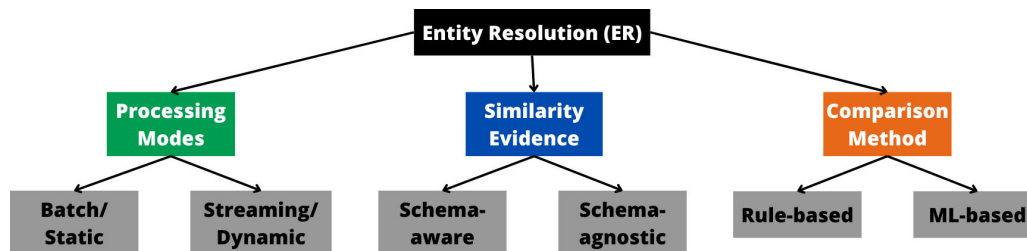


FIGURE 2. Taxonomy of ER approaches, extended from [1].

platforms or IoT devices [1], [23]. In such cases, schema-agnostic approaches are essential. These methods, which often utilize tokenization or embedding-based techniques, avoid schema dependencies by focusing on the inherent similarity between records. However, schema-agnostic ER introduces challenges, including increased computational complexity due to the larger feature space and potential reductions in accuracy when handling noisy or ambiguous data [28]. These challenges are exacerbated in streaming contexts, where schema-agnostic methods must process data incrementally without access to the entire dataset for comprehensive analysis.

Schema-aware techniques leverage the structural information and relationships defined within data schemas to enhance the matching process. These methods depend on predefined schemas and semantic mappings, which provide clarity on how data attributes correlate across different sources [29]. Such approaches are effective when schemas are consistent and well-defined, allowing algorithms to establish precise and structured comparisons, such as the approaches proposed in [29], [30], [31], [32], and [33].

Schema-agnostic techniques operate without prior knowledge of data schemas. They focus on the intrinsic content and statistical characteristics of the data, making them versatile for heterogeneous or semi-structured datasets where schema definitions are absent or highly variable [1]. Techniques such as blocking, clustering based on attribute similarities, and data-driven matching rules fall under this category. Works such [22], [25], [34], [35] are especially useful for scenarios involving large-scale, diverse datasets with unpredictable structures, as they avoid the overhead of schema alignment and can dynamically adapt to data variations. Also, Metablocking is an important family of schema-agnostic techniques known for its efficiency and effectiveness in reducing unnecessary comparisons [20], [36]. It restructures entity pairs into a weighted graph, applying pruning criteria to eliminate edges with low weights, thereby discarding unlikely comparisons [37].

How similar evidence is handled in ER significantly impacts challenges related to explanation and fairness. Schema-aware techniques leverage predefined attribute correspondences, making explanations more straightforward through feature attributions and explicit similarity metrics. However, they risk reinforcing biases if critical but non-standard attributes are excluded, potentially

marginalizing underrepresented groups. Ensuring fairness requires careful selection of attributes and fairness-aware similarity thresholds. On the other hand, schema-agnostic methods offer flexibility by relying on embedding-based or statistical approaches that do not require predefined schemas. While these methods enhance scalability and adaptability, they present challenges for explainability, as it becomes harder to trace how the similarity decisions are made. Additionally, fairness risks arise from latent biases in the data or embeddings, which may inadvertently favor majority groups. Addressing these challenges requires transparent embedding models and techniques that can highlight influential data patterns, even without explicit schema guidance.

C. RULE-BASED VS. ML-BASED COMPARISON METHODS

ER relies on comparison methods to evaluate record similarities and determine matches. Two primary approaches are rule-based and ML-based techniques. Rule-based approaches use manually defined similarity functions (e.g., Levenshtein distance, Jaccard similarity) and threshold-based decision rules. They provide high interpretability but require domain expertise and struggle with scalability in complex datasets. ML-based approaches leverage supervised and unsupervised models, such as decision trees, deep learning, and clustering algorithms, to automatically detect entity matches. These methods improve adaptability and scalability, but their closed-box nature makes explanation and fairness challenging. While rule-based methods ensure transparency, ML-based techniques enhance automation and accuracy. The choice between them depends on data complexity, domain requirements, and the need for interpretability in ER applications.

The integration of Artificial Intelligence (AI) in ER has significantly advanced the field, particularly in handling large-scale and heterogeneous data [6], [7]. AI-based methods, especially deep learning and pre-trained language models, have enhanced ER by overcoming limitations in traditional rule-based and statistical comparison methods (see Figure 2) [37], [38]. These models excel in capturing syntactic and semantic relationships between entity attributes, utilizing neural networks to learn latent representations. Pre-trained models like BERT embed contextual information into dense vectors, enabling nuanced understanding of data variations, including typographical errors, abbreviations, and attribute structure differences [6], [8], [38], [39]. However, challenges

remain, particularly regarding the computational cost of embedding generation and matching in large datasets [38]. Additionally, these models struggle in schema-agnostic environments, where attribute relationships are undefined. Addressing these issues requires adaptive techniques capable of inferring schema correspondences or functioning independently of predefined structures [7].

Another significant challenge pertains to scalability. AutoBlock [40] and DeepBlocker [41], address this by efficiently identifying candidate pairs through learned blocking schemes. However, these methods present trade-offs between recall and precision, and their effectiveness can vary depending on the input data [37]. Another issue is domain adaptation and generalizability, as pre-trained models often underperform in domains different from their original training data. While fine-tuning and domain-specific embeddings can improve results, they rely on labeled data, which may be scarce [8], [42]. This highlights the need for unsupervised or semi-supervised approaches that leverage unlabeled data to enhance adaptability. Finally, error propagation is a significant concern. Errors in early stages, such as blocking, can cascade and degrade the quality of final matches [22], [38]. Addressing this requires robust architectures that mitigate the impact of early-stage inaccuracies in the ER process.

Recent advancements in ER have leveraged deep learning and pre-trained language models to improve accuracy and scalability. DeepMatcher [9] introduced a framework evaluating various neural architectures, including SIF, RNNs, attention-based, and hybrid models, demonstrating that deep learning significantly outperforms traditional similarity-based methods. This work underscored the influence of architecture choices on performance and laid the groundwork for subsequent neural-based ER studies. Also, GNEM [42] proposed a graph-based approach that transitions from traditional pairwise matching to a one-to-set paradigm. By constructing record pair graphs and leveraging Graph Neural Networks (GNNs), GNEM captures contextual relationships, enhancing scalability and generalization, particularly in complex many-to-one and one-to-many matching scenarios. Incorporating Transformer-based models, Ditto [8] works on pre-trained models like BERT, RoBERTa, and DistilBERT for ER tasks, achieving state-of-the-art performance with minimal training data. Ditto also employs data augmentation to enhance robustness across domains. Expanding on Ditto's findings, [39] analyzed 12 pre-trained language models across 17 benchmark datasets, investigating schema-aware and schema-agnostic ER. Their results highlighted that Transformer-based models consistently outperform traditional embeddings but vary in effectiveness depending on dataset characteristics.

D. DISCUSSION

Summarizing the key contributions, methodologies, and characteristics of the reviewed works, the comparative analysis in Table 1 presents a structured overview of these

studies. It categorizes them based on their processing mode (batch or streaming), similarity evidence (schema-aware or schema-agnostic), and comparison method (rule-based or ML-based), offering insights into their respective approaches and applications in ER.

Even with significant advances in ER, challenges remain, particularly in fairness-aware and explainable ER for *batch vs. streaming* data. Most existing ER solutions focus on batch processing, where models have access to complete datasets. However, real-world applications increasingly demand streaming ER methods capable of handling continuously evolving data while maintaining fairness and transparency. The lack of approaches explicitly designed for streaming settings highlights a key research gap.

Regarding *schema-aware vs. schema-agnostic* approaches, schema-aware methods offer clearer explanations and structured attribute alignments, but they reinforce biases when schemas are misaligned or designed without fairness considerations. Schema-agnostic methods, on the other hand, provide greater flexibility, especially in heterogeneous datasets, but pose challenges in explainability and fairness monitoring, as hidden biases in embeddings or feature representations may lead to unintended discrimination. Future research should investigate how schema-agnostic models can integrate fairness-aware constraints and how schema-aware techniques can become more adaptable without rigid schema dependencies.

Similarly, *rule-based vs. ML-based* comparison methods introduce distinct challenges. Rule-based methods provide interpretability auditing, as decisions are based on explicit rules, but they could struggle with scalability and domain adaptability [2]. In contrast, ML-based approaches can capture complex relationships but often act as closed-box models, making fairness evaluation and bias detection more difficult. Future research should focus on hybrid models that balance interpretability with the adaptability of ML, incorporating XAI techniques to make closed-box models more transparent while ensuring bias mitigation mechanisms at each stage of the ER pipeline.

IV. UNDERSTANDING FAIRNESS AND EXPLAINABILITY

A. AN OVERVIEW OF FAIRNESS NOTIONS

Fairness in algorithmic systems is often framed as the absence of discrimination [13], [44], requiring that an algorithm's outputs are not unjustly influenced by input attributes irrelevant to the task. These attributes, known as protected or sensitive, commonly include gender, religion, age, and sexual orientation. Most existing work on fairness has centered on classification tasks, where each input entity is assigned to one of several predefined categories.

Fairness is commonly formalized through two perspectives [45]: (i) **Individual fairness**, where similar entities should receive similar treatment, and (ii) **Group fairness**, where entities grouped by protected attributes should be treated comparably. In individual fairness, similarity is often

TABLE 1. Comparative analysis of ER related works.

Work	Processing Mode	Similarity Evidence	Comparison Method
[29]–[33]	Batch	Schema-aware	Rule-based
[22]–[37], [43]	Batch	Schema-agnostic	Rule-based
[8], [9], [39]–[42]	Batch	Schema-aware	ML-based
[34]	Batch	Schema-agnostic	ML-based
[25]–[27]	Streaming	Schema-agnostic	Rule-based

captured via a task-specific distance metric $d : V \times V \rightarrow \mathbb{R}$ over the set of entities V , reflecting how alike two entities are. This metric may be domain-specific and should ideally reflect ground truth. Whether externally defined or proposed by stakeholders, the metric should be transparent and open to scrutiny. For group fairness, entities are divided into groups based on protected attributes. Care must be taken to consider proxy features—attributes correlated with protected ones—which can introduce redundant encoding.

Determining what constitutes similar treatment involves both social and technical considerations. Socially, it involves the distinction between [46]: (i) Equality, that is the uniform treatment regardless of need, and (ii) Equity, that is the adjusted treatment to ensure fair outcomes, accounting for existing disadvantages. Further social-based distinctions include [47]: (i) Disparate treatment, directly using protected attributes in decision-making, and (ii) Disparate impact, where outcomes disproportionately affect the protected groups, even without explicit use of sensitive attributes. Technically, the definition of fairness depends on the algorithm. For example, for rankings, that is part of a typical entity resolution workflow, a key concern is position bias—the tendency for top-ranked items to receive disproportionately more attention. Fair rankings aim to offer similar visibility to similar entities or groups of entities.

Next, we further refine individual and group fairness based on how similarities are specified.

1) TYPES OF INDIVIDUAL FAIRNESS

Individual fairness can be understood in several ways, with one prominent formulation being based on similarity metrics. The central idea is that if two individuals are similar according to some distance function d , then the algorithm's outputs for them should also be close, as measured by a corresponding output distance function D . This principle has been particularly influential in the context of probabilistic classifiers [45]. Another interpretation of individual fairness is known as counterfactual fairness [48]. Here, the fairness of a decision is evaluated by comparing the actual outcome to what would have happened in a hypothetical scenario where the individual had a different group membership. This approach relies on causal inference techniques to rigorously define and assess fairness across such alternate realities.

2) TYPES OF GROUP FAIRNESS

Assume two groups: a protected group G^+ and a non-protected group G^- . Let Y be the true label, \hat{Y} the

predicted label from a binary classifier, and S the predicted probability of the positive class (i.e., a favorable outcome). Then, group fairness definitions in classification can be broadly categorized into three families [49], [50]:

- *Base rate approaches*, using only predictions \hat{Y} ;
- *Accuracy-based approaches*, involving both \hat{Y} and Y ;
- *Calibration-based approaches*, relying on S and Y .

Base rate fairness assesses whether $P(\hat{Y} = 1 | v \in G^+)$ and $P(\hat{Y} = 1 | v \in G^-)$ are close, using either ratio [47], [51] or difference [52]. When these probabilities are equal, the condition is known as *statistical parity* or *demographic parity*, ensuring equal access to favorable outcomes across groups. *Accuracy-based fairness* requires that error rates, such as true or false positives, be similar across groups. A prominent example is *equal opportunity* [53], where the true positive rates are equal: $P(\hat{Y} = 1 | Y = 1, v \in G^+) = P(\hat{Y} = 1 | Y = 1, v \in G^-)$. *Calibration fairness* considers probabilistic predictions. A classifier is calibrated if, among individuals assigned score p , about p fraction truly belong to the positive class. Fair calibration demands this property holds across groups: $P(Y = 1 | S = p, v \in G^+) = P(Y = 1 | S = p, v \in G^-)$ [54], [55].

B. AN OVERVIEW OF EXPLAINABILITY NOTIONS

The widespread adoption of Large Language Models (LLMs) to address a broad range of tasks has generated renewed interest in the need for explainability, which seeks to provide human-understandable justifications or descriptions of model behavior. Due to the complexity and scale of recent AI models, achieving meaningful explainability remains a significant technical and conceptual challenge. Nonetheless, an increasing body of research [56] is dedicated to this issue, typically distinguishing between two main, high-level approaches: **Local explanations**, which aim to clarify individual model decisions or outputs, and **global explanations**, which seek to characterize the model's behavior or logic more generally. Finally, there is one more category that recently appeared in the bibliography, that of **glocal explanations** which try to combine local with global explanations.

Local explanations aim to provide an understanding of how a model makes a prediction for a specific input instance. This could be done with several slightly different approaches, namely: (i) **Feature attribution-based explanations**, which can further be divided into perturbation-based, gradient-based, decomposition-based or surrogate models. LIME [57] is a widely used method that belongs to the latter paradigm. (ii) **Attention-based explanations** [58], [59] try to identify

the most important parts of the input that affect the returned results, as a way of explaining a model's output. (iii) **Example-based explanations** include the widely adopted counterfactual explanations [60], [61], [62], as well as adversarial examples and data influence. Counterfactual explanations reveal what would have happened if certain input changes were observed (typically, as small changes as possible). A common example for counterfactually explaining the rejection decision of a loan application is that the application would have been approved if the personal income was higher by x amount. (iv) **Natural language explanations** [63] provide insights in a generated text in natural language regarding a decision taken by a model.

Global explanations aim to provide a broader understanding of how a model works, in general, by analyzing not only one input instance, but the global behavior of the model. Global explanation methods can be divided into (i) **probing methods** [64] that analyze model representations and parameters, and (ii) **neuron activation explanations** [65] that look for neurons that are the most important for model performance.

Finally, Global explanations aim to combine benefits from both local and global explanations. This can be applied for example by iteratively aggregating local explanations [66], or by introducing Shapley values, which are considered local explanations, on different levels of the model, making model-level generalizations and explanations [67]. However, this family of explanation methods is largely unexplored.

V. FAIRNESS IN ENTITY RESOLUTION

As the adoption of AI-driven systems increases, ethical considerations in their development and deployment have become an important area of study. Within the realm of ER, fairness has emerged as an important concern, particularly given the potential societal impacts of biased decision-making processes. ER systems are increasingly utilized in high-stakes applications, such as healthcare, e-commerce, and law enforcement, where fairness breaches can exacerbate systemic inequalities [6], [68].

Fairness in ER involves mitigating biases that arise during the whole process. These biases often stem from inherent imbalances in the datasets used to train ER algorithms. For instance, demographic overrepresentation or underrepresentation can lead to discriminatory outcomes, such as unequal error rates across protected and non-protected groups [69]. Additionally, biases can emerge from the design of algorithms themselves, as decision-making processes may inadvertently prioritize majority groups due to optimization criteria focused solely on accuracy [15].

The challenges of addressing fairness in ER are wide-ranging. Group fairness metrics, such as statistical parity and equal opportunity, aim to ensure equitable treatment across demographic groups. However, optimizing for these metrics can conflict with accuracy or individual fairness measures, where each entity is expected to be treated equitably based

on its unique characteristics. This challenge highlights the need for ER systems to balance competing fairness objectives, often necessitating trade-offs between ethical and accuracy goals [18], [70]. Another research topic in ER is related to fairness-aware ranking methods, which adjust similarity scores to account for demographic disparities in the input data. By incorporating fairness metrics directly into the ranking process, these methods aim to minimize disparate impacts during the resolution phase. Techniques such as proportional fairness constraints further refine these approaches, ensuring that entities from underrepresented groups are adequately represented in the final matching outputs [7], [15], [68].

FairER [15] was the first work to introduce fairness constraints into the matching process to mitigate bias and ensure equitable outcomes. Traditional ER systems optimize for accuracy but often overlook how bias in datasets and similarity measures can disproportionately affect specific groups. To address this, the authors integrate fairness constraints directly into the ER pipeline, ensuring that the resolution process does not systematically favor certain entity groups over others. The framework incorporates fairness constraints by adjusting similarity thresholds and modifying the matching process to minimize disparities while maintaining high accuracy. Their results highlight that introducing fairness constraints can significantly improve group-level equity without sacrificing overall performance.

Recent advancements in fairness-aware matching algorithms have introduced novel techniques to ensure equitable and diverse selections in streaming and online environments. In [70], the authors propose a framework for online matching with proportional fairness and diversity constraints, aiming to balance both the efficiency of data matching and fairness across different groups. The study formalizes fairness-aware matching as an optimization problem over hypergraphs, ensuring that selected matches adhere to predefined proportional representation criteria. Their approach introduces new approximate algorithms that efficiently enforce fairness constraints without significantly compromising computational efficiency. By extending traditional online matching frameworks with fairness considerations, this work paves the way for more equitable decision-making in real-time ER scenarios, where fairness is critical in domains such as job matching, recommendation systems, and dynamic resource allocation. In this perspective, the work [71] focuses on max-min diversity maximization in streaming and sliding-window settings, addressing challenges where fairness constraints must be continuously maintained as new data arrives. The study develops novel approximate algorithms to identify a diverse and fair subset of elements from an evolving data stream while considering sensitive attributes. Their results demonstrate that fairness-aware algorithms significantly outperform traditional selection methods in streaming environments, where data drift and evolving entity distributions pose major challenges.

The rise of streaming data has introduced additional complexities to fairness-aware ER. In streaming contexts, where data arrives incrementally, traditional batch processing methods fail to adapt to the dynamic and temporal nature of the input [72]. Despite recent advances, ensuring fairness in streaming ER remains an open challenge. One critical issue is the propagation of biases over time, as historical data inconsistencies may influence future decisions. To minimize, fairness-aware streaming algorithms incorporate mechanisms for bias detection and correction, leveraging historical data to refine decision-making processes [6], [7], [27].

Inserting fairness and ER into the streaming context, TREATS [7] is proposed as a fairness-aware ER approach for streaming data, addressing both the technical and ethical challenges associated with incremental ER. Unlike traditional ER approaches that focus primarily on accuracy, the proposed framework introduces fairness constraints, ensuring that matches are made in a way that reduces bias across different demographic groups. This is particularly relevant in real-time applications, where decisions must be made incrementally while maintaining fairness and efficiency. By tackling the intersection of streaming, fairness, and ER, TREATS advances by bridging gaps in fairness-aware ER methodologies, particularly in streaming environments where traditional batch approaches fall short.

Summarizing the main contributions, the comparative analysis in Table 2 provides a structured overview of the studies addressing fairness in ER, highlighting their proposed solutions, challenges, and impact on bias mitigation. Next, we also highlight insights and opportunities regarding how fairness can be considered in each step of ER.

A. BLOCKING

Blocking is designed to enhance efficiency by reducing the number of record comparisons while maintaining high effectiveness in ER approaches. Traditional blocking techniques rely on heuristic rules or learned representations to group similar records, yet these methods may accidentally introduce biases that disproportionately affect specific demographic groups or data distributions. Ensuring fairness at this stage is critical, as biases introduced during blocking can propagate throughout the ER pipeline, leading to systematic disparities in ER outcomes.

A primary source of bias in blocking arises from the selection of blocking keys, which determine how records are grouped before comparison. Many existing methods rely on attributes such as names, locations, or product categories, which can inadvertently exclude or over-represent certain subpopulations. For instance, demographic attributes can vary significantly across different cultural or regional contexts, and if these variations are not accounted for, blocking schemes may disproportionately filter out records from underrepresented groups. As a result, subsequent matching steps may suffer from lower recall for these groups, reinforcing existing disparities in entity linkage tasks.

Another challenge is the trade-off between recall and precision in fairness-aware blocking. Traditional approaches optimize blocking for computational efficiency and accuracy, yet fairness-aware methods must balance these objectives with equitable representation across groups. Techniques such as fairness-aware blocking adjust blocking keys dynamically to ensure that protected attributes, such as race, gender, or socioeconomic indicators, do not introduce systematic exclusion of specific groups. One potential approach is to use adaptive blocking mechanisms that incorporate fairness constraints, ensuring that records from diverse groups are proportionally distributed across blocks. However, such methods may introduce additional computational overhead, requiring careful design to maintain efficiency in large-scale or streaming ER settings.

Despite some recent advances, fairness-aware blocking remains an open research challenge. Future work should focus on developing scalable, fairness-preserving blocking techniques that can operate efficiently in both batch and streaming ER contexts. This includes exploring hybrid approaches that integrate fairness constraints while maintaining computational feasibility, as well as developing standardized fairness metrics tailored to the unique challenges of the blocking step. Addressing these gaps will be critical in ensuring that ER approaches remain both effective and equitable in their decision-making processes.

B. MATCHING

The matching step compares records using similarity measures, machine learning models, or hybrid techniques to generate match scores or classification decisions. However, biases present in the data, similarity functions, or model training processes can lead to unfair matching outcomes, disproportionately affecting certain groups. Addressing fairness in matching is fundamental, as disparities introduced at this step may result in systematic misclassification of entities, reinforcing social and economic biases in critical applications such as hiring, financial services, and law enforcement [6].

One of the primary sources of bias in matching comes from the similarity metrics used to compare entity attributes. Traditional similarity functions (e.g., Jaccard similarity, Levenshtein distance) and machine learning-based models often rely on textual, numerical, or categorical attribute comparisons that may inadvertently introduce bias. For example, matching based on the *name* attributes can exhibit cultural biases, favoring certain naming conventions over others, leading to lower accuracy for underrepresented groups. Similarly, product matching in e-commerce may favor well-known brands over local or minority-owned businesses due to imbalances in available training data. To mitigate these effects, fairness-aware similarity functions and embedding-based representations should be designed to account for variations in entity attributes across different demographic groups.

Machine learning-based matching models further introduce algorithmic biases when trained on historical datasets

TABLE 2. Comparative analysis of ER and fairness related works.

Work	Processing Mode	Similarity Evidence	Comparison Method
FairER [15]	Batch	Schema-aware	ML-based
Proportionally Fair Matching [70], Fair Max–Min [71]	Streaming	Schema-aware	Rule-based
TREATS [7]	Streaming	Schema-aware	ML-based

that reflect existing societal disparities. Supervised models often inherit biases from training data that may be skewed towards majority groups, leading to disproportionately higher false positive or false negative rates for underrepresented entities. Fairness-aware training strategies, such as re-weighting techniques, adversarial debiasing, and fairness constraints, can be integrated into matching models to ensure more equitable classification outcomes [13], [18]. Additionally, active learning approaches can be employed to iteratively refine model predictions, prioritizing the correction of biased match decisions during training.

In streaming ER, fairness challenges are compounded by the evolving nature of data. As new entity descriptions arrive over time, model drift can exacerbate disparities, particularly if updates to matching models disproportionately favor recently observed entity distributions. Fairness-aware incremental learning techniques can help mitigate these effects by continuously adapting models to maintain fair representations of different groups. One potential approach is to implement dynamic fairness constraints that periodically assess and correct for emerging biases in matching decisions.

Therefore, future research should focus on scalable fairness-aware matching techniques, particularly those capable of handling large-scale data while maintaining fairness constraints. Additionally, developing standardized fairness evaluation metrics designed for ER applications will be fundamental for assessing and improving fairness across diverse datasets. By addressing these challenges, fairness-aware matching can contribute to more ethical, reliable, and socially responsible matching algorithms across various domains.

C. CLUSTERING

Unlike blocking and matching, which focus on pairwise comparisons, clustering goes beyond direct matches to infer indirect relationships between records, enabling transitive closure. However, fairness concerns in clustering remain underexplored, even though biases introduced in earlier ER steps can be amplified or mitigated during this stage. Ensuring fairness in clustering is critical, as errors at this level can disproportionately impact certain groups, leading to biased representations in applications such as recommendation systems, hiring platforms, and financial risk assessments.

A major challenge in fairness-aware clustering is the propagation of biases from the matching stage. If the matching model systematically underrepresents certain groups (e.g., minority demographics or low-resource entities), clustering algorithms may fail to correctly consolidate their records,

fragmenting their representations across multiple clusters. This can result in higher false negatives for underrepresented groups while favoring well-connected, frequently occurring entities. Additionally, clustering methods that rely on pairwise confidence scores may reinforce existing biases if the similarity measures are not fairness-aware.

Another issue arises from *over-merging* and *over-splitting* trade-off. In *over-merging*, distinct entities may be grouped into the same cluster due to biased similarity metrics, leading to identity conflation (e.g., mistakenly grouping multiple individuals from the same ethnic background as one). On the other hand, *over-splitting* occurs when entity records are incorrectly divided into multiple clusters, often due to incomplete attribute similarities, which disproportionately affect marginalized groups whose data may be more fragmented across sources.

To mitigate these biases, fairness-aware clustering strategies can be introduced. One approach is fairness-constrained clustering, where clusters maintain demographic proportionality, preventing the over-representation or under-representation of specific groups [18]. Another technique involves reweighting similarity scores to reduce bias propagation from earlier ER steps, ensuring that underrepresented entities are not penalized. Additionally, group-aware clustering refines transitive closure methods by integrating diversity-aware constraints, rather than relying solely on confidence scores from matching [13].

To consolidate the discussion on fairness in entity resolution, it is important to highlight how fairness-aware interventions can be systematically integrated across the different stages of the ER pipeline. As explored in this section, each stage (i.e., blocking, matching, and clustering) offers distinct opportunities for addressing bias and enhancing equitable results. Blocking can incorporate bias-aware criteria, group-conscious pruning, and fair candidate generation to ensure diverse and representative candidate sets. Matching benefits from fairness-aware similarity learning, bias mitigation in embeddings, and fairness adjustments. Clustering can apply group-aware merge strategies and constraint-based methods to preserve fairness in the final resolved entities. Figure 3 visually summarizes these intervention points, illustrating how fairness mechanisms can be embedded throughout the ER workflow to support the development of socially responsible and trustworthy resolution systems.

VI. EXPLANATIONS IN ENTITY RESOLUTION

Explanation has emerged as a promising strategy for enabling transparency, trust, and accountability in AI-driven systems, particularly in the domain of data integration [18].

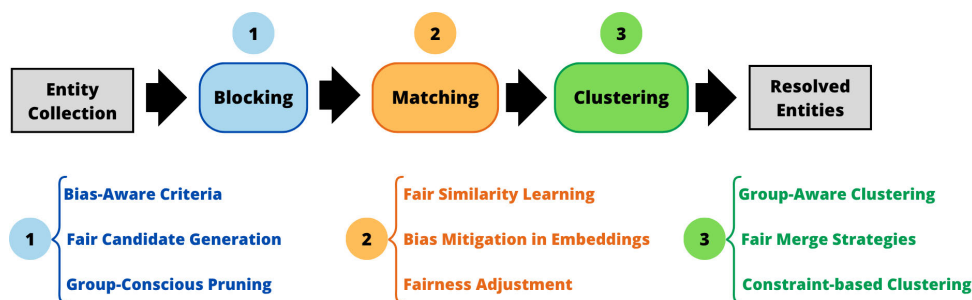


FIGURE 3. Fairness-aware interventions over ER workflow.

Data integration often involves consolidating heterogeneous and large-scale datasets to extract meaningful insights or align entities across diverse sources. The opacity of embedding-based and deep learning models, however, presents significant challenges in understanding and verifying their matching decisions. Specifically, research in XAI addresses these concerns by offering interpretability mechanisms tailored to the complexities of data integration processes [11], [16], [62].

The scalability and high-dimensionality of data integration tasks further complicate the application of XAI techniques [73]. Traditional methods, such as SHAP [74], often falter when faced with the computational demands of large-scale datasets. To address these limitations, novel approaches such as DALE [75] leverage auto-differentiation to efficiently estimate feature effects through Accumulated Local Effects (ALE) strategies, making them particularly suitable for data integration scenarios where explanations must scale with data volume and complexity. However, ALE strategies do not account for the heterogeneity of instance-level effects, which can obscure meaningful variations in feature importance. Additionally, it relies on a user-defined binning process that can introduce inconsistencies in estimations. In this sense, the RHALE method [76] overcomes these issues by introducing a novel approach to quantify heterogeneity through the standard deviation of local effects while optimizing the binning process to ensure unbiased estimation. By automatically determining variable-sized bins, RHALE balances bias and variance in feature effect estimations.

In [57], the authors propose LIME, a widely recognized explanation method. LIME aims to address the closed-box nature of complex models by approximating their local behavior with a simpler, interpretable model. The method achieves this by perturbing the input data, observing the resulting predictions, and then fitting a locally interpretable model, such as linear regression, to capture the relationship between input features and the model output. One of the strengths of LIME is its model-agnostic nature, meaning it can be applied to any classifier, regardless of the underlying algorithm. This characteristic allows it to be used across various domains, including text classification, image recognition, and recommendation systems.

ER exemplifies the utility of XAI in providing actionable insights for data integration. For instance, the authors of [11] introduce a framework for explaining and repairing embedding-based entity matching. Their work focuses on generating semantic matching subgraphs, which highlight the key relationships influencing entity matching decisions. The framework also includes repair mechanisms, allowing practitioners to adjust model decisions by providing insights into potential misalignments. A main contribution is its ability to uncover hidden biases in embeddings, demonstrating how interpretability techniques can be leveraged to improve model reliability. Similarly, inspired in LIME, the method LEMON [77] proposes a dual-explanation approach that combines feature attribution and counterfactual analysis. This method enhances the interpretability of embedding-based ER models by explaining why a pair of entities was matched while also presenting counterfactual scenarios where a different decision would have been made. LEMON emphasizes faithfulness and usability, ensuring that explanations accurately reflect the internal logic of deep learning models.

A critical challenge in deploying XAI in data integration is managing the trade-off between granularity and comprehensibility of explanations. While regional (or local) explanations provide detailed insights into specific decisions, they can overwhelm users, particularly when analyzing complex relationships in high-dimensional data. On the other hand, coarse (or global) explanations risk omitting essential details necessary for error correction or bias detection but avoid high computational costs. Expanding beyond local feature explanations, the framework Effector [12] focuses on regional explanations, enabling users to analyze entity matching at different granularity levels. This approach is particularly valuable in cases where global trends in entity matching need to be interpreted, rather than just individual pairwise decisions. Effector introduces novel techniques to define explanation regions, making complex ER models more accessible to data specialists. These studies demonstrate a shift towards explainable entity matching, bridging the gap between high-performing deep learning models and the need for transparent, human-interpretable decision-making processes in ER.

The need for counterfactual explanations has also gained traction, particularly in scenarios involving fairness and transparency. Counterfactual methods propose minimal changes to input data that could alter a model prediction, thus providing a pathway for users to understand and potentially influence outcomes. These techniques are fundamental to addressing biases inherent in data integration processes, such as systemic overrepresentation or underrepresentation of specific entity categories [11], [16], [62].

The integration of explanation methods in data integration offers both technical advancements and ethical improvements, fostering trust, transparency, and accountability in AI-driven systems. By elucidating decision-making processes, XAI enhances human-AI collaboration, ensuring that AI systems complement rather than replace human expertise [17]. Particularly in tasks such as quality control and anomaly detection, explanation methods has been shown to increase confidence in AI outputs by making reasoning more transparent [16]. In the context of ER, explanation provides mechanisms to identify and mitigate biases in matching decisions, supporting fair and reliable outcomes in applications such as credit scoring and recruitment [62]. By offering local explanations for embedding-based entity matching, explanation bridges the gap between data complexity and interpretability, enabling users to better understand model decisions without sacrificing accuracy [11]. The success of graph-based explanations in recommender systems suggests promising directions for enhancing explanation in relational data structures used in data integration and ER workflows [78]. Additionally, fairness-aware XAI frameworks help embed bias detection and correction mechanisms directly into ER workflows, ensuring equitable representation across different demographic groups [6], [7], [15].

In Table 3, we provide an overview of explanation methods in ER. The table classifies works based on *Domain* (domain-agnostic or specific), *Explanation Scope* (local and/or global), and *Explanation Technique* (e.g., feature importance, rule extraction, subgraph-based methods). This categorization highlights key strengths, limitations, and research gaps, offering insights into how different approaches improve explanation methods in ER.

In this work, we address the concept of eXplainable Entity Resolution (XER), which represents the integration of XAI techniques into the ER process. XER aims to enhance the transparency, interpretability, and trustworthiness of ER systems, particularly in complex environments characterized by heterogeneous data, streaming inputs, and fairness constraints [16]. As stated, traditional ER workflows involve several steps, such as data blocking, matching, and clustering, where decisions are often made by complex machine learning models that function as closed-boxes [6]. While these models achieve high accuracy, they lack mechanisms to explain how and why specific entity matches or non-matches occur. However, as ER systems increasingly influence decision-making processes in critical domains (such as healthcare, finance,

social networks, and e-commerce) the need for explanation becomes fundamental [7].

Explanations in ER, explored through counterfactual reasoning, feature attribution, and regional explanations, bridge the gap between closed-box models and interpretable decision-making. Therefore, XER aims to explain not only *what* decisions are made (e.g., whether two entities match) but also *why* those decisions were made. This involves generating human-interpretable insights into the factors that influenced entity matching, such as the contribution of specific attributes, the impact of model parameters, or the role of historical data in incremental processing scenarios. Moreover, XER facilitates the detection of biases and inconsistencies in ER models, supporting the development of fairness-aware approaches that are both effective and ethically responsible. In this sense, we would like to highlight insights and opportunities regarding how XAI can be applied in each step of ER.

A. BLOCKING

Blocking is a critical step in ER designed to improve computational efficiency by reducing the number of record pairs that need to be compared [2]. Traditional blocking methods rely on heuristic rules to group similar records, but these rules often lack transparency, making it difficult to understand why certain records are included or excluded from candidate sets [3], [79].

XAI techniques can enhance the interpretability of blocking decisions by providing clear explanations of the criteria used to form blocks. For example, feature attribution methods can identify which attributes (e.g., name, address, product ID) were most influential in determining block membership. Additionally, counterfactual explanations can be employed to explore how slight changes in data could alter blocking outcomes, revealing the sensitivity of the process to specific features or threshold values [62]. Moreover, in dynamic environments where data is continuously updated, such as streaming ER scenarios, explainable blocking models can provide real-time feedback on how new data points are integrated into existing blocks. This helps detect and correct blocking errors early in the pipeline, reducing the risk of matching inaccuracies [71].

B. MATCHING

The comparison phase involves measuring the similarity between pairs of records within the same block. This phase typically relies on similarity functions or machine learning models to generate comparison scores. However, the idea behind these similarity assessments is often unclear, especially when complex models like deep neural networks are used [6], [39].

XAI can be applied to clarify the comparison process by elucidating how similarity scores are derived. Techniques such as SHAP [74] and LIME [57] can decompose comparison scores into contributions from individual features,

TABLE 3. Comparative analysis of ER and XAI related works.

Work	Domain	Explanation Scope	Explanation Technique
Effector [12]	Domain-Agnostic	Local	Rule-based Regional Explanations
LIME [57]	Domain-Agnostic	Local	Feature Importance with Perturbation
DALE [75], RHALE [76]	Domain-Agnostic	Local and Global	Attribution-Based Explanations
LEMON [77]	Entity Matching	Local and Global	Logical Rule Extraction
ExEA [11]	Data Alignment	Local and Global	Subgraph-based and Dependency Graph

helping users understand which attributes drive high or low similarity [12]. Additionally, visual explanations, such as heatmaps or saliency maps, can highlight the specific parts of text or numerical data that most influence similarity judgments. In cases where multiple comparison functions are aggregated, the explanation can reveal the interaction between different similarity metrics. For example, if different similarity functions are used to determine comparison scores, XAI can reveal how each function contributes to the final comparison score, enabling better calibration of similarity thresholds and reducing false positives or negatives.

C. CLUSTERING

The final step of the ER workflow is clustering (or classification), where comparison scores determine whether record pairs represent the same entity in the real world. Classification models range from simple rule-based systems to complex ML-based classifiers. Regardless of the model type, XAI can significantly benefit the decision-making process.

In classification, XAI provides transparency into how features and comparison scores are weighted in the final decision. For instance, decision trees and ensemble models like random forests can be complemented with feature importance plots that highlight the most influential factors in matching decisions [11]. For closed-box models, such as neural networks, surrogate models and local explanations like LIME [57], SHAP [74], Effector [12], and LEMON [77] could approximate decision boundaries and provide interpretable insights. Furthermore, XAI supports fairness auditing in classification decisions. By analyzing the contribution of sensitive attributes (e.g., gender, ethnicity) to matching outcomes, XAI helps identify potential biases and informs the development of fairness-aware ER models [13], [15]. This is particularly important in applications with legal or ethical implications, such as credit scoring or recruitment [6].

In conclusion, XER represents a paradigm shift in how ER systems are designed and evaluated. By embedding explanation across all stages of the ER workflow, XER enhances transparency, fosters trust, and promotes fairness. This holistic approach not only improves the technical robustness of ER systems but also aligns them with broader societal expectations for ethical and accountable AI.

To synthesize the perspectives discussed throughout this section, Figure 4 illustrates how explanation techniques can be integrated across the ER workflow. In the blocking

phase, approaches such as feature attribution and interpretable blocking rules help clarify which attributes influence candidate generation, while pruning bias detection helps in identifying potentially unfair filtering. During the matching step, explanation methods like local and global explanations, attribute importance analysis, and counterfactual insights enhance the transparency of similarity computations and final match decisions. Finally, in the clustering phase, techniques including global clustering visualization, structured explanation models, and fairness auditing offer interpretability over merge strategies and final resolution outputs. These explanation-aware interventions not only provide clarity into the decision-making process at each stage but also serve as critical tools for identifying and mitigating potential biases in ER workflows.

VII. SUMMARY AND OPEN CHALLENGES

The increasing complexity and volume of data streams have elevated the significance of ER as a fundamental component of data integration processes across various domains [23]. In this paper, we explored the intersection of ER, XAI, and fairness, particularly within the context of streaming data environments. The analysis of state-of-the-art methods, combined with the introduction of the XER paradigm, provided insights into current capabilities, limitations, and open challenges that need to be addressed to advance the field. ER has evolved significantly from traditional rule-based approaches to more sophisticated machine learning techniques that leverage deep learning and pre-trained embeddings. However, the shift towards more complex models has introduced challenges related to transparency and interpretability. XAI emerges as a critical component to mitigate these challenges, enabling a better understanding of the logic behind the model's predictions.

The adoption of XAI in streaming ER workflows presents promising opportunities but also raises technical hurdles. The real-time nature of streaming data requires explanation mechanisms that can operate with minimal latency while providing informative and actionable insights. Current XAI techniques, have demonstrated potential for enhancing interpretability, but their scalability and efficiency in dynamic streaming environments remain areas for further exploration. Future research should investigate lightweight, adaptive, and context-aware explanation models tailored specifically for incremental ER tasks. The fairness dimension in ER also emerged as a fundamental concern, especially when dealing with datasets that reflect societal biases. However,

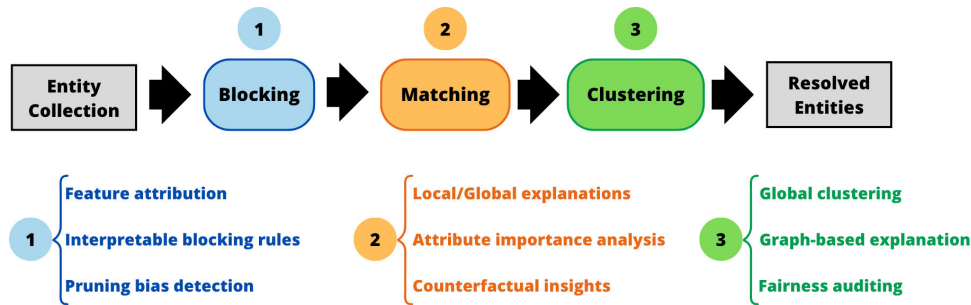


FIGURE 4. Explanations over ER workflow.

fairness-aware ER is still novel, particularly in streaming contexts where evolving data distributions can perpetuate or amplify biases over time. Addressing these concerns requires the development of bias detection and correction mechanisms that adapt to dynamic data patterns, ensuring equitable outcomes across diverse demographic groups [7].

From a practical perspective, the deployment of XAI-enabled ER systems has the potential to impact a wide range of applications, including fraud detection, healthcare analytics, e-commerce product matching, and social media monitoring. In these domains, decisions made by ER systems influence critical processes and outcomes, making the interpretability and fairness of these systems imperative [16], [17]. Looking ahead, several avenues for future research emerge from the insights presented in this work.

In this sense, XER opens up a range of research opportunities in the field of data integration, particularly in the development of more transparent, robust, and fair AI systems. Based on the current state of the art, several promising research directions emerge that can shape the future of XAI in ER.

One key opportunity lies in the development of **real-time explanation frameworks for streaming ER systems**. While existing XAI methods, such as SHAP and LIME, offer insights into model behavior, they are computationally intensive and not optimized for streaming contexts where decisions must be made incrementally and under strict latency constraints [71]. Research could focus on designing lightweight, adaptive explanation models that can operate in conjunction with ER systems, providing on-the-fly interpretability without compromising efficiency. This includes the exploration of approximation techniques, such as incremental feature attribution methods, that balance the trade-off between explanation granularity and computational efficiency.

Despite their applicability, existing fairness metrics in ER still face challenges when applied to **high-dimensional datasets** [80], [81]. These datasets often exhibit sparse or highly correlated attributes, which can lead to unreliable group definitions and unstable fairness measurements. Moreover, traditional fairness metrics such as statistical parity or equal opportunity may not capture nuanced biases introduced

during blocking or matching in complex ER workflows. Future work should investigate fairness metrics that account for structural properties of high-dimensional ER settings and are robust to attribute interactions and distribution shifts.

Another promising direction is the **integration of fairness-aware explanations in ER workflows**. Although fairness metrics have been explored in isolation, their connection with explanation remains unexplored. There is an opportunity to investigate how XAI can be leveraged not just to detect bias, but to actively guide fairness interventions during the ER process. For example, counterfactual explanations could be adapted to highlight not only what changes would alter a match decision, but also whether those changes introduce or mitigate bias [7], [62]. This could lead to the development of bias-aware explanation models that are sensitive to fairness constraints and capable of supporting ethical decisions in high-risk applications.

The **fusion of graph-based explanation techniques with ER models** represents another fruitful area for research. Given that many ER problems can be naturally modeled as graphs (e.g., entity linkage across networks), explainable graph neural networks (GNNs) could provide novel insights into how relationships between entities influence matching decisions [78]. Research could explore methods to generate interpretable subgraph explanations that reveal the structural dependencies driving model predictions. This is particularly relevant for applications like social network analysis, knowledge graph alignment, and fraud detection, where relational data plays a central role.

Moreover, there is an untapped potential in developing **domain-adaptive XAI techniques for ER**. Current explanation models often struggle to generalize across domains, especially when there are significant differences in data schemas, attribute semantics, or entity distributions [11]. This raises the need for research into transferable explanation mechanisms that can adapt explanations to new domains with minimal retraining. Techniques such as meta-learning for XAI or zero-shot explanation generation could be explored to enhance the robustness and scalability of ER in diverse environments.

Another critical opportunity involves addressing the **interpretability of composite ER pipelines**, where decisions are

influenced by multiple steps, including blocking, comparison, and classification. Traditional XAI approaches tend to focus on isolated models, neglecting the cumulative effects of the entire pipeline. There is a need to develop holistic explanation frameworks that provide end-to-end transparency, enabling users to trace errors or biases back to specific stages of the ER process [6], [15]. This could involve designing pipeline-aware attribution methods or multi-level explanation architectures that capture both local and global model behaviors.

Finally, the intersection of **human-centered XAI with ER systems** presents an attractive topic for research. While technical explanations are valuable, their effectiveness often depends on the user's ability to interpret and act on them. This highlights the importance of user-adaptive explanations that cater to different levels of expertise, from data scientists to non-technical stakeholders. Research could investigate interactive XAI tools that allow users to explore explanations dynamically, adjust model parameters based on feedback, and co-create knowledge with the system. The integration of visualization techniques, natural language explanations, and explainable dashboards could significantly enhance the accessibility and usability of ER systems [16], [17].

In conclusion, XER is not only a complement to interpretability, but a changing way that can redefine how ER systems are designed, evaluated, and deployed. Therefore, this paper has provided a comprehensive overview of the current state of ER, the integration of AI techniques, the growing relevance of fairness, and the transformative potential of XAI. By introducing the XER paradigm, the paper outlines a path toward ER systems that are more transparent, equitable, and adaptable to the dynamic nature of modern data streams. The challenges discussed herein present fertile ground for future research, with the potential to shape the development of next-generation ER systems that not only achieve high performance but also support principles of transparency, fairness, and trustworthiness in data-driven decision-making.

REFERENCES

- [1] V. Christophides, V. Efthymiou, T. Palpanas, G. Papadakis, and K. Stefanidis, "An overview of end-to-end entity resolution for big data," *ACM Comput. Surveys*, vol. 53, no. 6, pp. 1–42, Nov. 2021.
- [2] P. Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Cham, Switzerland: Springer, 2012.
- [3] V. Christophides, V. Efthymiou, and K. Stefanidis, "Entity resolution in the Web of data," *Synth. Lectures Semantic Web*, vol. 5, no. 3, pp. 1–122, 2015.
- [4] N. Ayat, R. Akbarinia, H. Afsarmanesh, and P. Valduriez, "Entity resolution for probabilistic data," *Inf. Sci.*, vol. 277, pp. 492–511, Sep. 2014.
- [5] X.-L. Liu, H.-Z. Wang, J.-Z. Li, and H. Gao, "EntityManager: Managing dirty data based on entity resolution," *J. Comput. Sci. Technol.*, vol. 32, no. 3, pp. 644–662, May 2017.
- [6] N. Shahbazi, N. Danevski, F. Nargesian, A. Asudeh, and D. Srivastava, "Through the fairness lens: Experimental analysis and evaluation of entity matching," *Proc. VLDB Endowment*, vol. 16, no. 11, pp. 3279–3292, Jul. 2023.
- [7] T. B. Araújo, V. Efthymiou, V. Christophides, E. Pitoura, and K. Stefanidis, "Treats: Fairness-aware entity resolution over streaming data," *Inf. Syst.*, vol. 129, Mar. 2025, Art. no. 102506.
- [8] Y. Li, J. Li, Y. Suhara, A. Doan, and W.-C. Tan, "Deep entity matching with pre-trained language models," *Proc. VLDB Endowment*, vol. 14, no. 1, pp. 50–60, Sep. 2020.
- [9] S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, and V. Raghavendra, "Deep learning for entity matching: A design space exploration," in *Proc. Int. Conf. Manage. Data*, May 2018, pp. 19–34.
- [10] K. Ma and B. Yang, "Stream-based live entity resolution approach with adaptive duplicate count strategy," *Int. J. Web Grid Services*, vol. 13, no. 3, pp. 351–373, 2017.
- [11] X. Tian, Z. Sun, and W. Hu, "Generating explanations to understand and repair embedding-based entity alignment," in *Proc. IEEE 40th Int. Conf. Data Eng. (ICDE)*, May 2024, pp. 2205–2217.
- [12] V. Gkolemis, C. Diou, E. Ntoutsis, T. Dalamagas, B. Bischl, J. Herbinger, and G. Casalicchio, "Effector: A Python package for regional explanations," 2024, *arXiv:2404.02629*.
- [13] E. Pitoura, K. Stefanidis, and G. Koutrika, "Fairness in rankings and recommendations: An overview," *VLDB J.*, vol. 31, no. 3, pp. 1–28, May 2022.
- [14] K. Makhlof, S. Zhioua, and C. Palamidessi, "Machine learning fairness notions: Bridging the gap with real-world applications," 2020, *arXiv:2006.16745*.
- [15] V. Efthymiou, K. Stefanidis, E. Pitoura, and V. Christophides, "FairER: Entity resolution with fairness constraints," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2021, pp. 3004–3008.
- [16] S. Moosavi, M. Farajzadeh-Zanjani, R. Razavi-Far, V. Palade, and M. Saif, "Explainable AI in manufacturing and industrial cyber-physical systems: A survey," *Electronics*, vol. 13, no. 17, p. 3497, Sep. 2024.
- [17] K. Nikiforidis, A. Kyrtoglou, T. Vafeiadis, T. Kotsiopoulos, A. Nizamis, D. Ioannidis, K. Votis, D. Tzovaras, and P. Sarigiannidis, "Enhancing transparency and trust in AI-powered manufacturing: A survey of explainable AI (XAI) applications in smart manufacturing in the era of industry 4.0/5.0," *ICT Exp.*, vol. 11, no. 1, pp. 135–148, Feb. 2025.
- [18] C. Fragkathoulas, V. Papanikou, D. P. Karidi, and E. Pitoura, "On explaining unfairness: An overview," in *Proc. IEEE 40th Int. Conf. Data Eng. Workshops (ICDEW)*, May 2024, pp. 226–236.
- [19] G. Papadakis, G. Alexiou, G. Papastefanatos, and G. Koutrika, "Schema-agnostic vs schema-based configurations for blocking methods on homogeneous data," *Proc. VLDB Endowment*, vol. 9, no. 4, pp. 312–323, Dec. 2015.
- [20] G. Simonini, L. Gagliardelli, S. Bergamaschi, and H. V. Jagadish, "Scaling entity resolution: A loosely schema-aware approach," *Inf. Syst.*, vol. 83, pp. 145–165, Jul. 2019.
- [21] B.-H. Li, Y. Liu, A.-M. Zhang, W.-H. Wang, and S. Wan, "A survey on blocking technology of entity resolution," *J. Comput. Sci. Technol.*, vol. 35, no. 4, pp. 769–793, Jul. 2020.
- [22] G. Papadakis, V. Efthymiou, E. Thanos, O. Hassanzadeh, and P. Christen, "An analysis of one-to-one matching algorithms for entity resolution," *VLDB J.*, vol. 32, no. 6, pp. 1369–1400, Nov. 2023.
- [23] K. Nikolettos, E. Ioannou, and G. Papadakis, "The five generations of entity resolution on Web data," in *Proc. Int. Conf. Web Eng.*, 2024, pp. 469–473.
- [24] Y. Wang, F. Fabbri, and M. Mathioudakis, "Streaming algorithms for diversity maximization with fairness constraints," in *Proc. IEEE 38th Int. Conf. Data Eng. (ICDE)*, May 2022, pp. 41–53.
- [25] T. B. Araújo, K. Stefanidis, C. E. S. Pires, J. Nummenmaa, and T. P. da Nóbrega, "Incremental entity blocking over heterogeneous streaming data," *Information*, vol. 13, no. 12, p. 568, Dec. 2022.
- [26] D. Karapiperis, C. Tjortjis, and V. S. Verykios, "A randomized blocking structure for streaming record linkage," *Proc. VLDB Endowment*, vol. 16, no. 11, pp. 2783–2791, Jul. 2023.
- [27] D. Karapiperis, C. Tjortjis, and V. S. Verykios, "A suite of efficient randomized algorithms for streaming record linkage," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 7, pp. 2803–2813, Jul. 2024.
- [28] T. B. Araújo, C. E. S. Pires, D. G. Mestre, T. P. D. Nóbrega, D. C. D. Nascimento, and K. Stefanidis, "A noise tolerant and schema-agnostic blocking technique for entity resolution," in *Proc. 34th ACM/SIGAPP Symp. Appl. Comput.*, Apr. 2019, pp. 422–430.

- [29] D. G. Mestre, C. E. S. Pires, D. C. Nascimento, A. R. M. de Queiroz, V. B. Santos, and T. B. Araújo, "An efficient spark-based adaptive windowing for entity matching," *J. Syst. Softw.*, vol. 128, pp. 1–10, Jun. 2017.
- [30] D. Paulsen, Y. Govind, and A. Doan, "Sparkly: A simple yet surprisingly strong TF/IDF blocker for entity matching," *Proc. VLDB Endowment*, vol. 16, no. 6, pp. 1507–1519, Feb. 2023.
- [31] B. Ramadan, P. Christen, H. Liang, and R. W. Gayler, "Dynamic sorted neighborhood indexing for real-time entity resolution," *J. Data Inf. Qual.*, vol. 6, no. 4, pp. 1–29, Oct. 2015.
- [32] P. Rosso, D. Yang, N. Ostapuk, and P. Cudré-Mauroux, "RETA: A schema-aware, end-to-end solution for instance completion in knowledge graphs," in *Proc. Web Conf.*, Apr. 2021, pp. 845–856.
- [33] J. Wang, Y. Li, and W. Hirota, "Machamp: A generalized entity matching benchmark," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, 2021, pp. 4633–4642.
- [34] K.-S. Teong, L.-K. Soon, and T. T. Su, "Schema-agnostic entity matching using pre-trained language models," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2020, pp. 2241–2244.
- [35] G. Simonini, S. Bergamaschi, and H. V. Jagadish, "BLAST: A loosely schema-aware meta-blocking approach for entity resolution," *Proc. VLDB Endowment*, vol. 9, no. 12, pp. 1173–1184, Aug. 2016.
- [36] G. Papadakis, G. Papastefanatos, T. Palpanas, and M. Koubarakis, "Scaling entity resolution to large, heterogeneous data with enhanced meta-blocking," in *Proc. 19th Int. Conf. Extending Database Technol. (EDBT)*, Bordeaux, France, E. Pitoura et al., Eds., OpenProceedings.org, Mar. 2016, pp. 221–232, doi: [10.5441/002/EDBT.2016.22](https://doi.org/10.5441/002/EDBT.2016.22).
- [37] G. Papadakis, D. Skoutas, E. Thanos, and T. Palpanas, "Blocking and filtering techniques for entity resolution: A survey," *ACM Comput. Surveys*, vol. 53, no. 2, pp. 1–42, Mar. 2021.
- [38] A. Zeakis, G. Papadakis, D. Skoutas, and M. Koubarakis, "Pre-trained embeddings for entity resolution: An experimental analysis," *Proc. VLDB Endowment*, vol. 16, no. 9, pp. 2225–2238, May 2023.
- [39] A. Zeakis, G. Papadakis, D. Skoutas, and M. Koubarakis, "An in-depth analysis of pre-trained embeddings for entity resolution," *VLDB J.*, vol. 34, no. 1, pp. 1–27, Jan. 2025.
- [40] W. Zhang, H. Wei, B. Sisman, X. L. Dong, C. Faloutsos, and D. Page, "AutoBlock: A hands-off blocking framework for entity matching," in *Proc. 13th Int. Conf. Web Search Data Mining*, Jan. 2020, pp. 744–752.
- [41] D. Javdani, H. Rahmani, M. Allahgholi, and F. Karimkhani, "DeepBlock: A novel blocking approach for entity resolution using deep learning," in *Proc. 5th Int. Conf. Web Res. (ICWR)*, Apr. 2019, pp. 41–44.
- [42] R. Chen, Y. Shen, and D. Zhang, "GNEM: A generic one-to-set neural entity matching framework," in *Proc. Web Conf.*, Apr. 2021, pp. 1686–1694.
- [43] V. Efthymiou, G. Papadakis, K. Stefanidis, and V. Christophides, "MinoanER: Schema-agnostic, non-iterative, massively parallel resolution of Web entities," in *Proc. 22nd Int. Conf. Extending Database Technol.*, Lisbon, Portugal, Mar. 2019, pp. 373–384.
- [44] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making," *Commun. ACM*, vol. 64, no. 4, pp. 136–143, Apr. 2021.
- [45] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel, "Fairness through awareness," in *Innovations in Theoretical Computer Science*. Cambridge, MA, USA: ACM, 2012, pp. 214–226.
- [46] M. Minow, "Equality vs. equity," *Amer. J. Law Equality*, vol. 1, pp. 167–193, Sep. 2021, doi: [10.1162/ajle_a_00019](https://doi.org/10.1162/ajle_a_00019).
- [47] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Sydney, NSW, Australia, Aug. 2015, pp. 259–268.
- [48] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2017, pp. 4066–4076.
- [49] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," in *Proc. Conf. Fairness, Accountability, Transparency*, Atlanta, GA, USA, Jan. 2019, pp. 329–338.
- [50] S. Verma and J. Rubin, "Fairness definitions explained," in *Proc. Int. Workshop Softw. Fairness*, Gothenburg, Sweden, May 29, 2018, pp. 1–7.
- [51] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proc. 30th Int. Conf. Mach. Learn.*, Atlanta, GA, USA, Jun. 2013, pp. 325–333.
- [52] T. Calders and S. Verwer, "Three naive Bayes approaches for discrimination-free classification," *Data Mining Knowl. Discovery*, vol. 21, no. 2, pp. 277–292, Sep. 2010.
- [53] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proc. Adv. Neural Inf. Process. Syst. 29: Annu. Conf. Neural Inf. Process. Syst.*, Barcelona, Spain, 2016, pp. 3315–3323.
- [54] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, Jun. 2017.
- [55] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *Proc. 8th Innov. Theor. Comput. Sci. Conf.*, 2016, pp. 43:1–43:23.
- [56] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du, "Explainability for large language models: A survey," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 2, pp. 1–38, Apr. 2024, doi: [10.1145/3639372](https://doi.org/10.1145/3639372).
- [57] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.
- [58] O. Barkan, E. Houn, A. Caciularu, O. Katz, I. Malkiel, O. Armstrong, and N. Koenigstein, "Grad-SAM: Explaining transformers via gradient self-attention maps," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2021, pp. 2882–2887, doi: [10.1145/3459637.3482126](https://doi.org/10.1145/3459637.3482126).
- [59] C. Yeh, Y. Chen, A. Wu, C. Chen, F. B. Viégas, and M. Wattenberg, "Attentionviz: A global view of transformer attention," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 1, pp. 262–272, Jan. 2024, doi: [10.1109/TVCG.2023.3327163](https://doi.org/10.1109/TVCG.2023.3327163).
- [60] T. Wu, M. T. Ribeiro, J. Heer, and D. Weld, "Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 6707–6723, doi: [10.18653/v1/2021.acl-long.523](https://doi.org/10.18653/v1/2021.acl-long.523).
- [61] D. Kaushik, E. Hovy, and Z. C. Lipton, "Learning the difference that makes a difference with counterfactually-augmented data," in *Proc. 8th Int. Conf. Learn. Represent.*, Apr. 2019, pp. 1–6.
- [62] X. Wang, Q. Li, D. Yu, Q. Li, and G. Xu, "Counterfactual explanation for fairness in recommendation," *ACM Trans. Inf. Syst.*, vol. 42, no. 4, pp. 1–30, Jul. 2024.
- [63] N. F. Rajani, B. McCann, C. Xiong, and R. Socher, "Explain yourself! Leveraging language models for commonsense reasoning," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, Jul. 2019, pp. 4932–4942, doi: [10.18653/v1/p19-1487](https://doi.org/10.18653/v1/p19-1487).
- [64] B. Chen, Y. Fu, G. Xu, P. Xie, C. Tan, M. Chen, and L. Jing, "Probing BERT in hyperbolic spaces," in *Proc. 9th Int. Conf. Learn. Represent. (ICLR)*, Virtual Event, Austria, OpenReview.net, May 2021. [Online]. Available: <https://openreview.net/forum?id=17VnwXYZyH>
- [65] F. Dalvi, N. Durrani, H. Sajjad, Y. Belinkov, A. Bau, and J. Glass, "What is one grain of sand in the desert? Analyzing individual neurons in deep NLP models," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jan. 2019, pp. 6309–6317, doi: [10.1609/aaai.v33i01.33016309](https://doi.org/10.1609/aaai.v33i01.33016309).
- [66] M. Setzu, R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti, "GLocalX—from local to global explanations of black box AI models," *Artif. Intell.*, vol. 294, May 2021, Art. no. 103457, doi: [10.1016/j.artint.2021.103457](https://doi.org/10.1016/j.artint.2021.103457).
- [67] E. Borghonovo, E. Plischke, and G. Rabitti, "The many Shapley values for explainable artificial intelligence: A sensitivity analysis perspective," *Eur. J. Oper. Res.*, vol. 318, no. 3, pp. 911–926, Nov. 2024, doi: [10.1016/j.ejor.2024.06.023](https://doi.org/10.1016/j.ejor.2024.06.023).
- [68] Y. Wang, W. Ma, M. Zhang, Y. Liu, and S. Ma, "A survey on the fairness of recommender systems," *ACM Trans. Inf. Syst.*, vol. 41, no. 3, pp. 1–43, Jul. 2023.
- [69] Z. Wang, N. Saxena, T. Yu, S. Karki, T. Zetty, I. Haque, S. Zhou, D. Kc, I. Stockwell, X. Wang, A. Bifet, and W. Zhang, "Preventing discriminatory decision-making in evolving data streams," in *Proc. ACM Conf. Fairness Accountability Transparency*, Jun. 2023, pp. 149–159.
- [70] A. Louis, M. Nasre, P. Nimbhorkar, and G. S. Sankar, "Online algorithms for matchings with proportional fairness constraints and diversity constraints," in *Proc. ECAI*, 2023, pp. 1601–1608.
- [71] Y. Wang, F. Fabbri, M. Mathioudakis, and J. Li, "Fair max–min diversity maximization in streaming and sliding-window models," *Entropy*, vol. 25, no. 7, p. 1066, Jul. 2023.

- [72] T. B. Araújo, K. Stefanidis, C. E. S. Pires, J. Nummenmaa, and T. P. da Nóbrega, "Schema-agnostic blocking for streaming data," in *Proc. 35th Annu. ACM Symp. Appl. Comput.*, Mar. 2020, pp. 412–419.
- [73] G. Giannopoulos, G. Papastefanatos, D. Sacharidis, and K. Stefanidis, "Interactivity, fairness and explanations in recommendations," in *Adjunct Proc. 29th ACM Conf. User Model., Adaptation Personalization*, Jun. 2021, pp. 157–161.
- [74] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowl. Inf. Syst.*, vol. 41, no. 3, pp. 647–665, Dec. 2014.
- [75] V. Gkolemis, T. Dalamagas, and C. Diou, "DALE: Differential accumulated local effects for efficient and accurate global explanations," in *Proc. Asian Conf. Mach. Learn.*, 2022, pp. 375–390.
- [76] V. Gkolemis, T. Dalamagas, E. Ntoutsis, and C. Diou, "RHALE: Robust and heterogeneity-aware accumulated local effects," in *Proc. ECAI*, 2023, pp. 859–866.
- [77] N. Barlaug, "Lemon: Explainable entity matching," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 8, pp. 8171–8184, Aug. 2022.
- [78] H.-M. Attolou, K. Tzompanaki, K. Stefanidis, and D. Kotzinos, "Why-not explainable graph recommender," in *Proc. IEEE 40th Int. Conf. Data Eng. (ICDE)*, May 2024, pp. 2245–2257.
- [79] G. Simonini, L. Zecchini, S. Bergamaschi, and F. Naumann, "Entity resolution on-demand," *Proc. VLDB Endowment*, vol. 15, no. 7, pp. 1506–1518, Mar. 2022.
- [80] J. R. Foulds, R. Islam, K. N. Keya, and S. Pan, "Bayesian modeling of intersectional fairness: The variance of bias," in *Proc. SIAM Int. Conf. Data Mining*, 2020, pp. 424–432.
- [81] M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro, "Bias mitigation for machine learning classifiers: A comprehensive survey," *ACM J. Responsible Comput.*, vol. 1, no. 2, pp. 1–52, Jun. 2024.



VASILIS EFTHYMIU received the Ph.D. degree in entity resolution in the Web of data from the Computer Science Department, University of Crete (UOC), Greece, in 2017. He is currently an Assistant Professor with the Department of Informatics and Telematics, Harokopio University of Athens (HUA), Greece. Before joining HUA, he was a Postdoctoral Researcher with the Database Group of IBM Research in Almaden Research Center, CA, USA, a Postdoctoral Researcher with the Information Systems Laboratory of FORTH-ICS, Greece, and a Visiting Instructor with UOC. After his Ph.D. research internship with IBM T. J. Watson Research Center, NY, USA, on matching Web tables to Knowledge Graphs (KGs), he has been co-organizing the SemTab challenges at ISWC, an effort to benchmark systems dealing with the tabular data to KG matching problem, and the Tabular Data Analysis (TaDA) workshop at VLDB. He has co-authored two books, more than 60 papers, and co-invented four US patents. He is currently serving as the Coordinator for the W3C Greek Office.



KOSTAS STEFANIDIS received the Ph.D. degree in personalized data management from the University of Ioannina, Greece. He is currently a Professor of data science with the Faculty of Information Technology and Communication Sciences, Tampere University, Finland, where he also leads the Data Science Research Centre and the Group on Recommender Systems. He has more than ten years of experience in different roles at ICS-FORTH, Greece, NTNU, Norway, and CUHK, Hong Kong. His work focuses on personalization and recommender systems, entity resolution, data exploration, and data analytics, with a special focus recently on socio-technical aspects in data management like fairness and transparency, and has been published in more than 100 papers in top-tier conferences and journals. He has been involved in several international and national research projects, and he is also actively serving the scientific community. His research interest includes the broader area of big data. He is the General Co-Chair of ADBIS 2025, TPD L 2025, and EDBT/ICDT 2026.

• • •



TIAGO BRASILEIRO ARAÚJO received the Ph.D. degree in computer science from UFCG, Finland. He is currently pursuing the second Ph.D. degree with Tampere University. He is currently a Professor with the Federal Institute of Paraíba (IFPB) and a Researcher with the VIRTUS Competence Center and the Applied Intelligent Computing Laboratory (LACINA), Federal University of Campina Grande (UFCG), Brazil. His Ph.D. research stay at Tampere University. He has co-authored papers in top-tier venues and actively collaborates on both research and industry-driven projects in AI-driven solutions, data analytics, and decision-support systems. His research interests include entity resolution, cloud computing, artificial intelligence, and data analytics, with a growing emphasis on the socio-technical aspects of data management, particularly fairness and transparency.