Towards Identifying Intent of Data Errors

Mohamed Abdelmaksoud
BIFOLD & TU Berlin
Berlin, Germany
mohamed@tu-berlin.de

Konrad Rieck
BIFOLD & TU Berlin
Berlin, Germany
rieck@tu-berlin.de

Ziawasch Abedjan
BIFOLD & TU Berlin
Berlin, Germany
abedjan@tu-berlin.de

ABSTRACT

Modern machine learning (ML) systems deployed in high-stakes domains such as hiring, lending, and healthcare heavily rely on structured, often user-provided input data. Errors in this data can arise from natural causes, such as noise, missing values, typos, or from strategic user manipulation intended to alter decision outcomes. Existing ML pipelines typically treat all input errors uniformly, lacking mechanisms to distinguish between accidental errors and intentional manipulations. The goal of this research is to develop a diagnostic framework that identifies erroneous input features, estimates the likelihood that each error was intentional, and quantifies its influence on the model's output. In this paper, we outline the foundational challenges of our research agenda. We discuss risks and potentials in trying to separate intentional from non-intentional errors.

VLDB Workshop Reference Format:
Mohamed Abdelmaksoud, Konrad Rieck, and Ziawasch Abedjan. Towards Identifying Intent of Data Errors. VLDB 2025 Workshop: VLDB Guide-AI Workshop.

1 INTRODUCTION

Modern machine learning (ML) systems are routinely deployed in high-stakes automated decision systems (ADS), such as hiring platforms, loan approvals, and healthcare triage, where predictions depend on structured, often user-provided input data [31]. Errors in this input are common and can arise from a variety of sources: from natural mistakes, e.g., typos, missing values, to deliberate attempts of outcome manipulation [16]. While there is a large body of work on error detection and correction [1, 19, 26], reasoning about the origin and intention of data errors has been neglected. Data errors may result from *unintentional* benign noise, distribution shifts, or user confusion [28, 29], or from *intentional* input manipulation.

Intentional data manipulation has been observed in collaborative knowledge bases, such as Wikipedia and Wikidata [32], where some users deliberately insert falsified or misleading content, known as vandalism, to manipulate public information or disrupt downstream systems. One notable form of such manipulation is coordinated disinformation campaigns, often executed by so-called *Web Brigades* [27]. Unlike unintentional user mistakes, these edits are often subtle, syntactically correct, and socially manipulative, designed to bypass detection systems. Recent advances in multilingual vandalism detection have shown that distinguishing such data corruptions requires careful modeling of both content semantics and user behavior, including fairness-aware evaluation to avoid bias against new or anonymous contributors [12, 32]. Yet, there has been little attention to this topic in common data curation pipelines, especially when structured datasets drive high-stakes decisions or automated learning processes.

Insight into user intent can inform safeguards during data collection and foster awareness of underlying socio-economic patterns. The lack of such knowledge undermines both trust and accountability. A strategically altered input feature, such as falsified education credentials in hiring or inflated income in a loan application, can lead to incorrect, unfair, or even dangerous decisions [31].

From a technical perspective, understanding the origin of data errors is essential for data quality assurance tasks; for instance, to design effective cleaning routines that target the most impactful and plausible corrections, ultimately guiding more responsible downstream decisions (e.g., which features to prioritize for repair or what repair procedure to choose [20]). For example, if an error is identified as intentional and shows signs of targeted manipulation, its correction should not merely revert it to a frequent value, but instead account for the likely direction of manipulation. Correcting an intentionally inflated income field should lead to a plausible lower value that neutralizes the manipulation's intended effect.

Diagnosing intent behind data errors remains fundamentally challenging, due to missing ground truth labels, ambiguous user behavior, context-sensitive manipulation. Thus, in this paper we argue for research on what drives certain errors to occur in the first place. Are they the result of deliberate manipulation, or simply random mistakes? We begin our analysis informed by ML security principles [6, 30] from the attack perspective: what kind of adversarial or strategic behavior could result in input manipulation, and what are the associated incentives? With this approach we transform the problem of distinguishing intentional and non-intentional data errors, which might be intractable, into the problem of identifying the occurrence of well-defined attacks.

Identifying *Intentional* errors generally requires the identification of an error and an intrinsic intent. Historically, there has been the awareness for intentional errors that aim at *system-level disruption*. Here, the adversaries intentionally introduce erroneous inputs with the goal of degrading the performance of a downstream model, corrupting decision pipelines, or undermining trust in the ML system as a whole (*Destroy Model Integrity*) [5, 6, 9, 18, 30]. This type of attack has received considerable attention by the research community in machine learning and we do not aim at expanding it further. Instead, we focus on intentional errors that aim at manipulating the implications of the data usage regarding individual data points or groups of data entries. Here we identify *Individual-level*

Proceedings of the VLDB Endowment. ISSN 2150-8097.

and *Group-level* manipulation as the two main categories that we will further define later.

Examples of *Individual-level manipulation* can include the following cases. A single user falsifies or omits information, e.g., inflating income or fabricating education credentials to increase their chance for a favorable decision. Another example is that minority users alter data inputs in response to a historically unfair system, e.g., masking sensitive attributes to avoid discriminatory outcomes [31]. A further form may emerge in the obfuscation of personal data using outlier-like placeholders, e.g., '1000' for age gain or '0' for blood pressure [24]. While such entries might reflect privacy-protecting behavior rather than malicious intent, they may still have strong model impact and warrant diagnostic attention.

In case of *Group-level manipulation*, data is strategically altered to improve outcomes for an entire group or to discriminate others under a particular data-specific policy or treatment. Such manipulation can take both coordinated and emergent forms. An illustrative example for the former is when a subgroup, e.g., women in a workplace, collectively adjust self-reported working hours to meet eligibility thresholds for promotion [31].

Research Questions: This work is motivated by several long-term research questions:

- How can we systematically characterize and distinguish patterns of intentional versus unintentional input errors in datasets that are subject to downstream tasks, such as ML pipelines? To what extent are defined attack models and the resulting input errors correlated with feature importance, outcome incentives, and systematic shifts in decision boundaries, and how effectively do these factors model intentional manipulation?
- What are efficient algorithms to capture heuristics that describe manipulated data and model slices? Is there a common pattern among different types of attacks or do we need individual detection algorithms for different types of intent? How dataset- or domain-specific is effective intent detection?
- What are the evaluation benchmarks and metrics suitable for distinguishing intentional and non-intentional errors?

Contributions: Our paper lays the groundwork for a broader research agenda on intent identification. Rather than presenting a finalized system, we articulate core challenges, formalize key concepts, and propose foundational components that guide our future research in this emerging space. As a step toward answering these questions, this paper contributes a foundational conceptual framework for intent attribution:

- We introduce a taxonomy of manipulation types, including falsification, masking, obfuscation, and coordinated group-level attacks, spanning both individual and collective behaviors.
- We define attacker capabilities and constraints, formalizing a threat model over tabular datasets.
- We propose a suite of interpretable heuristics to estimate manipulation likelihood and outcome significance. We outline a lightweight, adaptive aggregation scheme for combining these signals in a weakly-supervised way.

2 RELATED WORK

The challenges posed by anomalous or manipulated inputs have been explored from multiple angles across the machine learning literature, including adversarial robustness, error detection, fairness, and data quality. However, these research directions typically operate in isolation, each addressing a different slice of the broader input reliability problem. In this section, we review relevant work across these domains, highlighting both their strengths and their limitations in handling semantically meaningful, intent-driven input perturbations in structured data.

Adversarial Inputs to ML Systems. Adversarial inputs have been extensively studied in the context of unstructured data, particularly in image classification, where small, imperceptible perturbations can significantly alter model predictions. Seminal works such as the Fast Gradient Sign Method (FGSM) by Goodfellow et al. [9], Projected Gradient Descent (PGD) by Madry et al. [18], and the Carlini & Wagner (CW) attack [5] demonstrated how deep learning models are vulnerable to gradient-based adversarial perturbations. More recently, research has extended adversarial attacks to tabular and structured data domains. Cartella et al. [6] propose model-agnostic adversarial techniques for imbalanced tabular datasets in fraud detection scenarios. Their work revealed that even minimal but targeted changes in input features can mislead standard classifiers.

Complementing these works, He et al. [11] provide a reproducible benchmark for evaluating adversarial attacks on tabular data using both effectiveness and imperceptibility metrics. Their framework includes implementations of PGD, C&W, DeepFool [21], and LowProFool [3], and systematically quantifies perturbation impact through proximity, sparsity, and sensitivity scores. Pierazzi et al. [23] highlight key challenges in generating realistic attacks under semantic, syntactic, and contextual constraints, particularly when manipulating structured objects like PDFs or Android apps. Ghamizi et al. [8] introduce CoEvA2, a search-based adversarial testing framework for credit scoring systems under domain constraints. Their method employs multi-objective evolutionary algorithms to generate adversarial examples that optimize for model evasion, perturbation effort, and gain, while satisfying real-world feature constraints, e.g., monotonicity, value bounds. Building on the idea of realistic adversarial behavior in structured domains, Simonetto et al. [30] introduced CAA, a method for generating adversarial examples in constrained tabular deep learning models.

Trustworthy and Fair ML. Fairness in ML systems depends not only on model design but also on upstream processes, such as data collection and cleaning [31]. Schelter et al. [28] propose *FairPrep*, a framework that adjusts preprocessing to improve fairness, showing that routine cleaning operations can significantly shift fairness outcomes. However, such methods assume unintentional bias and overlook strategic user adaptations.

Recent work acknowledges that users may manipulate inputs to navigate biased systems, e.g., through résumé whitening or masking sensitive attributes [31]. Yet, fairness-enhancing pipelines treat all deviations alike, ignoring distinctions between noise, adaptation, and adversarial manipulation. Guha et al. [10] further caution that automated cleaning can worsen disparities, revealing that data quality interventions alone do not guarantee fairness.

Other efforts focus on identifying marginalized groups. Dehghankar and Asudeh [7] propose a method to discover underperforming cohorts without requiring sensitive attributes, but their approach targets structural disparities, not intentional manipulation.

Related studies on integrity threats in collaborative environments include vandalism detection. Heindorf et al. [12] and Trokhymovych et al. [32] propose fairness-aware techniques for detecting harmful edits in Wikidata and Wikipedia. While these works address user intent, they operate in crowd-sourced settings and focus on overt vandalism, not subtle manipulations in decision systems.

Error Detection in ML Pipelines. Error detection in ML systems is a well-studied area spanning adversarial robustness, data validation, and out-of-distribution (OOD) input handling. A common line of work focuses on identifying whether the model is likely to mispredict a given input, especially under adversarial conditions. *RED*, proposed by Qiu and Miikkulainen [25], proposes a residual-based detection system that uses Gaussian Process modeling to estimate the likelihood of misclassification. Their framework computes the residual between classifier confidence and true correctness, treating high-variance predictions as signals of potential OOD or adversarial inputs. While RED provides a flexible, model-agnostic approach to detect model uncertainty, it does not attempt to explain the *cause* of an error, whether it stems from benign noise, adversarial tampering, or user manipulation. Nor does it distinguish OOD from adversarial inputs in a structured diagnostic sense.

Bahat et al. [2] propose an image-specific adversarial detection method leveraging prediction instability under input transformations. Although effective in vision tasks, the approach does not generalize to structured tabular data, where manipulations are semantic rather than perceptual. CIAI, introduced by Jain et al. [15], proposes CIAI, a classifier for distinguishing between clean, noisy, and adversarially perturbed images using Vision Transformers. CIAI employs center loss and distribution-aware metrics to learn embeddings for three perturbation classes. While the notion of separating intentional from unintentional input corruption aligns conceptually with our goals, CIAI is confined to synthetically perturbed image data and assumes attacks generated by adversarial algorithms, e.g., FGSM, PGD. As a result, it does not generalize to semantically meaningful manipulations that occur in human-entered data or decision-critical applications like loan approvals or hiring. Orthogonal to adversarial detection, several systems target general error detection and data validation.

Error detection systems, such as Raha [19] and Matelda [1], focus on detecting inconsistent, missing, or anomalous records in structured datasets. They do not consider user intent, or the causal relationship between an input error and a model's decision.

Data Cleaning Prioritization. Data cleaning in ML pipelines is often resource-constrained, requiring selective intervention under fixed budget, or human supervision limits. This has led to a growing body of work on cleaning recommendation systems that aim to prioritize actions based on their estimated downstream utility. For example, Naumann et al. [20] propose Comet, a step-by-step data cleaning framework that estimates the predictive benefit of cleaning each feature and recommends a cleaning plan that maximizes model accuracy under a fixed budget. Similarly, HoloClean [26] performs holistic error repair using probabilistic inference over constraints, external data, and co-occurrence patterns, prioritizing cells whose correction is most statistically justified.

These methods typically assume that all detected errors are structurally similar, e.g., due to noise, missing values, or inconsistencies, and make no distinction between errors arising from accidental vs. intentional behavior. As a result, existing frameworks may allocate cleaning effort toward benign anomalies while overlooking manipulations introduced with adversarial or incentive-driven intent.

Summary of related work

While adversarial robustness, fairness auditing, and data quality research have addressed various forms of input anomalies, to the best of our knowledge, no existing framework systematically models the *intent* behind tabular data errors in machine learning pipelines. Current approaches typically focus on detecting prediction errors, identifying adversarial examples, or mitigating unfair outcomes, but treat data errors without distinguishing between unintentional errors and strategic manipulations.

In adversarial ML, the emphasis has largely been on constructing attacks or certifying robustness, not on interpreting the nature or motivation of real-world inputs once received [5, 8, 9, 18]. Similarly, fairness frameworks may flag bias-amplifying inputs or sensitive attribute usage, but do not assess whether such features were manipulated in response to systemic disparities [28, 29]. Even in data quality research, the focus is on correcting inconsistencies, not explaining why they occurred [20].

3 PROBLEM DEFINITION

Understanding whether errors in structured datasets are natural or intentional is critical for ensuring the integrity, fairness, and robustness of modern ML systems. This section introduces the core task of inferring the intent behind input errors. Further, it outlines threat scenarios and constraints that define the knowledge and capabilities of the incentivized actor (adversary).

3.1 Intent Attribution in Structured Data

The core objective is to distinguish between *intentional* and *unintentional* errors and assess their potential impact on downstream decision systems. To distinguish the two types of errors, we need an additional signal that serves as a proxy for the intent.

Problem Statement. Given a relational instance D, a set of detected errors $\mathcal{E}(D)$, and a model M that is applied on each instance of D, the objective is to assign each $e_i \in \mathcal{E}(D)$ a score $I(e_i)$ indicating how likely it is to be *intentional*:

$$I : \mathcal{E}(D) \rightarrow [0, 1]$$

A high score reflects great likelihood of strategic or adversarial manipulation, while low scores indicate lack of intention.

The ambiguity of intent is further compounded by the absence of explicit intent annotations in real-world datasets. Since user motivations are latent and rarely documented, traditional supervised learning approaches are generally infeasible. As a result, intent attribution must rely on indirect indicators, such as behavioral cues, causal impact, or rarity, rather than ground truth labels. Most causal analysis tools further assume a *ceteris paribus* condition, i.e., that a single feature can be changed in isolation while all other inputs remain fixed [22]. In structured data, features are often correlated, making such counterfactual evaluations potentially unrealistic [33]. The resulting estimates may overstate or understate the causal significance of an error if applied naively. Consequently, even well-intentioned corrections may have negligible or misleading downstream effects unless they are paired with a nuanced understanding of user motivation, domain context, and model behavior. Nevertheless, with the following taxonomy we try to identify subproblems that are tractable with appropriate heuristics.

3.2 Taxonomy for Manipulation Types

Figure 1 presents our taxonomy of intentional manipulations. The top-down layout distinguishes between *individual-level manipulation* and *group-level manipulation.* The former covers gain-targeted perturbations, fairness-triggered masking, information obfuscation, and individual discrimination. The latter includes collective manipulation and group discrimination. This taxonomy is not derived from a formal theory or exhaustive empirical enumeration. Rather, it reflects our current best attempt to systematize general and recurring manipulation patterns observed in tabular datasets used in ML decision systems. Its structure aims to support reasoning about the nature and scope of intent, guide the design of diagnostic heuristics, and provide realistic evaluation aligned with different manipulation types [4, 11, 13, 17, 31]. While the taxonomy captures the most common manipulation intentions observed in both adversarial and fairness-driven contexts, we acknowledge that completeness and uniqueness are inherently difficult to guarantee. Future work may uncover alternative behaviors or decompositions beyond this structure. *Next, we examine these manipulation types in more detail, beginning with individual-level intentions.*

3.2.1 Individual-level manipulation. The first category of intentional data manipulation from our taxonomy captures cases where a single user falsifies or omits input features to change their data-dependent treatment and cases where a third-party with full access to the dataset is trying to change the treatment of an individual. So far, we distinguish four subcategories. *Gain-targeted perturbation* covers cases where individuals choose properties that are not accurate and change their treatment positively. *Fairness-triggered masking* is a common approach by individuals from underrepresented groups that fear discrimination to blend in by sensitive properties, such as gender or ethnicity. Another type of protection is *information obfuscation*, which aims to preserve privacy. This is often the case for disguised missing values for personal information that a user does not want to share. Such obfuscation might appear as an outlier or as an inlier through disguised missing values [24]. Finally, it is also possible for a third party with access to individual records to discredit or favor specific individuals by manipulating their properties, representing a form of *individual discrimination.* For example, in collaborative data platforms, malicious insiders may intentionally alter performance ratings, or background information to harm a colleague [11]. These manipulations are independently affecting only a single data record.

Diagnosing whether an individual-level error is intentional is particularly challenging due to the semantic ambiguity of user behavior. A small change to a high-impact feature, such as increasing income by a few hundred dollars, might result from a simple typo, or a deliberate attempt to cross a decision threshold. Conversely, large and seemingly suspicious errors, that may or may not flip the predictions might stem from benign confusion, especially when users are unaware of how features are defined or used by the system. As such, neither the magnitude of an error nor its effect on the model's output solely is a sufficient indicator of intent.

To grasp the complexity in capturing the intent categories mentioned above, consider Table 1, where a snapshot of the Adult Income dataset [4] is shown. Also consider Table 2, where in the same snapshot some data points have been changed. Comparing both tables, changing the `Occupation` field for the instance at index 1 from `Protective-serv` to `Tech-support` flips the model prediction from ≤ $50K to > $50K, despite all other features remaining constant. Even a small error, as observable for the record with index 10, where two digits in the `CapGain` feature are swapped, is sufficient to flip the prediction outcome. Conversely, heavier perturbations did not change the model prediction. For instance, index 3 includes three coordinated modifications: upgrading the `Education` level to `Doctorate` and its corresponding numerical index `EdNum`, and changing the `Country` from `Mexico` to `United-States`.

Which of the three rows was manipulated intentionally is not obvious. While the first two both change the treatment, the swapped digits seem to resemble a typical human error in typing while the misclassification of the occupation might be more informed as in case of gain-targeted perturbation. In the last case, we do not observe a change in the treatment, which would be a signal against intention. Yet, we cannot exclude a manipulation as a mixture of by changing the education level that was in fact intentional, but not effective at changing the treatment.

3.2.2 Group-level manipulation. In contrast to individual manipulation, this category involves coordinated or correlated errors that affect the treatment of an entire group. In our taxonomy, we distinguish two cases. First, a group may deliberately coordinate its entries to influence the treatment of either their own group or another group. Second, a third-party with access to all records may deliberately change properties that affect a represented group. A group may be latent, defined by combinations of attribute ranges, e.g., Middle Eastern males aged 18–25.

A foundational challenge in attributing group-level intent is that group identities can be implicit or latent. Unlike protected attributes such as race or gender, which may be explicit, many strategic collectives must be inferred from behavioral regularities, shared input manipulations, or correlated downstream effects. This requires solving a non-trivial clustering problem over high-dimensional, often mixed-type data—balancing demographic proximity with coordination signals such as similar perturbation patterns or synchronized prediction shifts. Determining which users form a coherent group is computationally intensive and sensitive to the choice of similarity metrics, distance thresholds, and clustering algorithms. Moreover, there is an inherent risk of overgeneralization: coincidental resemblance across users may be mistaken for strategic coordination, especially in heterogeneous populations.

Diagnosing whether group-level errors are intentional is further challenging because intent must not be inferred from isolated anomalies, but from structured patterns across multiple records. However, not all group-level similarities imply coordinated manipulation: users from similar demographics or regions may coincidentally exhibit correlated errors due to shared social context, cultural

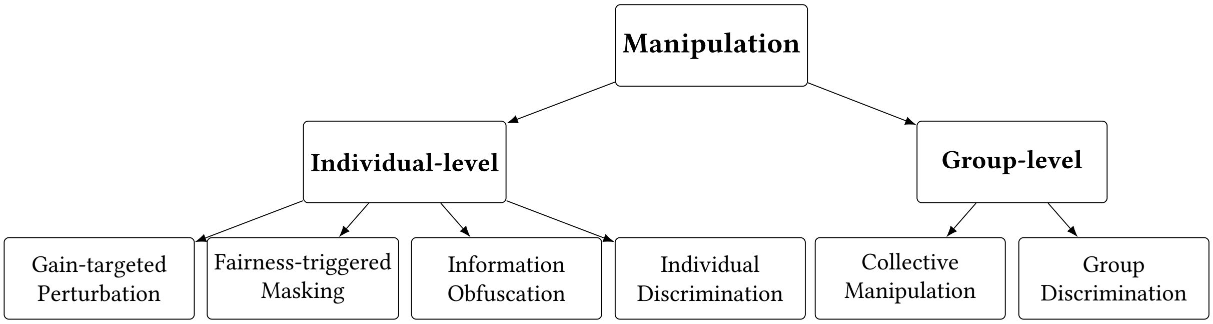

Figure 1: Taxonomy of Manipulation Types, distinguished by intent scope.

Table 1: Partial snapshot of the Adult Income dataset [4] containing complete and error-free feature values.

Index	Age	Workclass	Education	EdNum	Marital Status	Occupation	Race	Sex	CapGain	CapLoss	Hours	Country	Class
1	41	State-gov	Assoc-acdm	12	Married-civ-spouse	Protective-serv	White	Male	0	0	40	United-States	≤50K
2	24	Private	Masters	14	Never-married	Exec-managerial	White	Male	6849	0	90	United-States	≤50K
3	23	Private	HS-grad	9	Never-married	Other-service	White	Male	0	0	35	Mexico	≤50K
4	46	Private	Some-college	10	Married-civ-spouse	Tech-support	White	Male	0	0	40	United-States	>50K
5	31	Private	Bachelors	13	Married-civ-spouse	Prof-specialty	Asian-Pac-Islander	Male	15024	0	48	Philippines	>50K
6	39	Private	Masters	14	Married-civ-spouse	Prof-specialty	Asian-Pac-Islander	Male	0	0	40	Philippines	>50K
7	39	Private	Bachelors	13	Married-civ-spouse	Prof-specialty	Asian-Pac-Islander	Male	0	0	48	Philippines	>50K
8	39	Private	Bachelors	13	Married-civ-spouse	Craft-repair	Asian-Pac-Islander	Male	0	0	40	Philippines	>50K
9	46	Private	Prof-school	15	Married-civ-spouse	Prof-specialty	White	Male	99999	0	60	United-States	>50K
10	35	Self-emp-not-inc	9th	5	Married-civ-spouse	Craft-repair	White	Male	2635	0	30	United-States	≤50K

Note: These predictions are made using a Gradient Boosting Classifier (GBC) trained on the Adult dataset's training partition, achieving an F1-score of 87.7% on the held-out test set. The shown records in the above table are part of the test data from the same dataset.

Table 2: The snapshot from Table 1 with injected errors: erroneous values in bold violet, flipped labels in bold red. A group is marked between dotted lines.

Index	Age	Workclass	Education	EdNum	Marital Status	Occupation	Race	Sex	CapGain	CapLoss	Hours	Country	Class
1	41	State-gov	Assoc-acdm	12	Married-civ-spouse	**Tech-support**	White	Male	0	0	40	United-States	**>50K**
2	24	Private	Masters	14	Never-married	Exec-managerial	White	Male	6849	0	90	United-States	≤50K
3	23	Private	**Doctorate**	**16**	Never-married	Other-service	White	Male	0	0	35	**United-States**	≤50K
4	46	Private	Some-college	10	Married-civ-spouse	Tech-support	White	Male	0	0	40	United-States	>50K
5	31	Private	Bachelors	**5**	Married-civ-spouse	Prof-specialty	Asian-Pac-Islander	Male	**0**	0	48	Philippines	**≤50K**
6	39	Private	Masters	**5**	Married-civ-spouse	Prof-specialty	Asian-Pac-Islander	Male	0	0	40	Philippines	**≤50K**
7	39	Private	Bachelors	**5**	Married-civ-spouse	Prof-specialty	Asian-Pac-Islander	Male	0	0	48	Philippines	**≤50K**
8	39	Private	Bachelors	**5**	Married-civ-spouse	Craft-repair	Asian-Pac-Islander	Male	0	0	40	Philippines	**≤50K**
9	46	Private	Prof-school	15	Married-civ-spouse	Prof-specialty	White	Male	99999	0	60	United-States	>50K
10	35	Self-emp-not-inc	9th	5	Married-civ-spouse	Craft-repair	White	Male	**6235**	0	30	United-States	**>50K**

Note: These predictions are made using a Gradient Boosting Classifier (GBC) trained on the Adult dataset's training partition, achieving an F1-score of 87.7% on the held-out test set. The shown records in the above table are part of the test data from the same dataset.

patterns, or systemic biases. Distinguishing such organic group behavior from strategic group manipulation is non-trivial.

In our study using the Adult Income dataset, we observe in Table 1 that records in rows 5 to 8 form a cluster (confirmed via K-Nearest Neighbor analysis with $k = 4$) that can be characterized as *Married Philippine Males*. In Table 2, we see naive but impactful errors by decreasing the `Education-Num` without modifying the `Education` label for all four records, moreover changing the *CapGain* for the 5th record. These changes collectively shift the predicted outcomes for the entire group from > \$50K to ≤ \$50K, achieving the goal of group discrimination.

3.3 Operational Constraints

Effective intent attribution requires understanding not only what errors exist, but which of them were plausibly introduced by a strategic actor. In structured datasets, this depends on the manipulability of individual features, some of which are system-locked, while others are user-controlled and prone to falsification.

To ground our problem in real-world scenarios, we define a categorization of feature-level constraints that classifies attributes into three categories: immutable, softly immutable, and mutable. These categories shape both our simulation of adversarial behavior and the heuristics used in intent attribution.

Table 3 summarizes instance features from structured datasets, categorized by their manipulation feasibility as immutable, softly immutable, or mutable. *Immutable* features, such as tax IDs or birth dates, are system-controlled and rarely manipulable. Thus, errors in these fields are more likely to reflect unintentional pipeline issues. *Softly immutable* features, e.g., race, gender, or education, are editable but semantically constrained, hence strategic manipulation in these fields often signals fairness-sensitive adaptation. *Mutable*

Table 3: Examples of structured features and their categories.

Category	Feature	Manipulation Characteristics
Immutable	Name	Legally fixed; identity-verifiable
	Date of birth	Verified via documents; non-editable in formal systems
	Tax ID	Assigned by national authority; cannot be self-declared
	Credit score	Computed from upstream financial data; user cannot set directly
	Total debt	System-controlled; imported from verified financial sources
Softly Immutable	Race	Self-declared; manipulation may reflect socio-strategic intent
	Gender	Editable; sometimes omitted or misrepresented
	Age	Declared or derived; manipulation plausible in non-verified settings
	ZIP code	User-entered; verifiable through indirect features
	Education	Often unverifiable at entry point; falsification possible
Mutable	Income	Self-reported; high manipulation incentive to gain favorable outcomes
	Job title	Unverified, editable field; hard to validate without background check
	Work-hours	Manually declared; weakly enforced
	Willingness to relocate	Declarative; no validation or constraints applied

attributes, such as income or work hours, are user-declared and typically associated with strong incentives for falsification.

Our categorization bounds the adversarial space by identifying which features users can realistically manipulate, thereby highlighting regions where intent is diagnostically plausible. Even when limited to mutable features, adversaries may mimic benign changes to evade detection, motivating our formal modeling of their knowledge and capabilities.

3.4 Adversary Models in Intent Attribution

A common concern in attack scenarios is whether adversaries can manipulate their inputs strategically without triggering diagnostic intent signals. To this end, we consider *gray-box* and *white-box adversaries.* A *black-box adversary* is not considered here as data manipulation requires some insights into the usage of the data and we do not consider blind attacks targeting the system.

First, we assume a *gray-box adversary*, an attacker with partial knowledge of the model and dataset. Such adversary understands that extreme deviations or high-impact changes may be flagged by intent heuristics, but lacks access to internal scoring weights, precise heuristic definitions, or real-time outcomes. This scenario typically applies to individual users attempting to evade detection in automated decision systems, or to coordinated groups engaging in local manipulation.

Second, we acknowledge the presence of a *white-box adversary*, which is particularly relevant in settings such as federated learning or collaborative data pipelines, where some actors may have access to the global dataset and/or internal model. One instance of a white-box adversary in our setting is *group-level white-box adversary*, who can exploit known feature distributions or statistical artifacts to craft manipulations that evade detection while inducing collective shifts. Their deeper knowledge poses a more significant threat to diagnostic robustness and demands principled defenses across group-based intent attribution. Another variant is the *individual-level white-box adversary.* This type refers to an attacker with full knowledge of the system and the data, who focuses on manipulating a specific individual's record. In this case, the adversary can craft plausible errors that closely resemble benign noise, mirroring typical patterns, plausible ranges, and error distributions. This level of precision renders most diagnostic heuristics ineffective, approaching the theoretical limits of intent attribution.

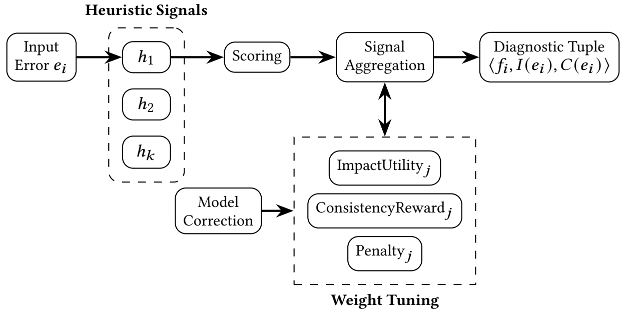

Figure 2: Overview of the proposed diagnostic framework. Heuristic signals are scored and aggregated into an intent score, producing a diagnostic tuple per error. Feedback from model correction informs utility-driven weight updates.

4 METHODOLOGY

Rather than seeking provable robustness, the goal of our research is to ensure that manipulations, across the spectrum defined in our taxonomy of attacks and scenarios (Figure 1), leave detectable traces in at least one signal. We ensure this through two design principles, aligned with the manipulation types captured in our taxonomy. First, we construct heuristics that are semantically and operationally diverse, spanning incentive-based, data-based, and effort-aware factors, so that evading detection across all signals simultaneously becomes increasingly difficult. Second, we will incorporate cross-heuristic consistency checks: for instance, a small perturbation with large causal effect triggers an effort-impact mismatch.

This section presents our proposed solution framework, outlines the evaluation plan, and the limitations of our approach.

4.1 Solution Concept

Building on the problem definition, our conceptual solution aims to compute a diagnostic tuple for each error $e_i \in \mathcal{E}(\mathcal{D})$:

$$\langle f_i, I(e_i), C(e_i), S(e_i) \rangle$$

Where f_i is the feature index in the input instance where the error occurred; $I(e_i) \in [0, 1]$ is the *intent score*, estimating the likelihood that the error is intentional rather than natural or benign; $C(e_i) \in [0, 1]$ is the *causal impact score*, quantifying how much the error e_i contributed to a change in the model prediction $f(x)$; $S(e_i)$ is an optional semantic or contextual classification of the error, e.g., “user manipulation”, “preprocessing bug”, “noise”. The diagnostic objective is to generate these tuples $\langle f_i, I, C, S \rangle$ independently of the model and without relying on ground truth labels.

Our framework in Figure 2 decomposes the intent estimation problem into two core components: (1) the design of interpretable heuristic signals $h_1, \ldots, h_k$ that each capture a distinct indicator of potential manipulation, and (2) adjusting the weights needed for aggregating these signals into a unified intent score $\mathcal{I}(e_i)$.

4.1.1 Heuristics. To achieve our goals, we first leverage a set of interpretable, modular factors that together capture the likelihood of intentional manipulation and the significance of the error on the model's output. The following heuristics are derived from recurring manipulation patterns observed in our taxonomy and are intended as a foundational set, expandable or adaptable depending on the manipulation scenario.

- *Feature Importance.* Features that strongly influence the model prediction are more likely targets for intentional manipulation.
- *User Incentive.* Certain features, e.g., income, education level, naturally incentivize manipulation due to their direct impact on outcomes like loan approvals. Incentive scores can be assigned based on domain knowledge or learned heuristics.
- *Causal Impact on Prediction.* We estimate how much altering the erroneous feature e_i would change the model's output. In other words, we compute counterfactual changes to model outputs to estimate the causal impact $C(e_i)$.
- *Error Rarity.* We assess how unusual the erroneous value is compared to the typical distribution of f_i in the dataset. While rarity alone does not imply intent, highly uncommon values may contribute to the suspicion of intentional manipulation when combined with other factors, such as outcome incentives or perturbation size.
- *Group Shift Potential.* We evaluate the extent to which the error changes the instance's position relative to decision boundaries or cohort-based groupings. This signal captures how much the manipulated feature shifts the input toward another outcome cluster, e.g., from a loan rejection group to an approval group, using clustering, classification confidence, or boundary proximity.
- *Minority-Sensitive Feature Handling.* Errors involving protected or socially sensitive attributes, e.g., race, gender, disability status, are flagged for special consideration, as they may reflect strategic adaptations to avoid discrimination. To automatically estimate the sensitivity of a feature, we envision leveraging large language models (LLMs) to score the social or ethical salience of feature names or descriptions, e.g., using prompt-based querying or embedding similarity to known fairness-sensitive concepts.
- *Similarity to True Value.* We assess how close the erroneous value is to its corrected value, using task-appropriate similarity measures. For example, small character-level changes may indicate typographical mistakes in text fields, while minimal numeric deviations may suggest entry slips rather than strategic manipulation.
- *Consistency.* If there are duplicate representations available in the data, we compare f_i to those entries for the same user or similar users in a group. Large unexpected deviations increase the likelihood of intent. Moreover, intra-group comparisons might represent good signal for detecting group-level intentional manipulation, i.e are similar users in the same group jointly deviating in a feature?
- *Effort-Based Perturbation Size.* We measure the minimal perturbation needed to achieve another outcome. Strategic errors often show minimal but significant changes. Each factor outputs an interpretable signal, which is then aggregated, via weighted sum to compute the final intent score $I(e_i)$.

Designing and operationalizing these heuristics poses several conceptual and practical challenges. First, each heuristic must transform abstract notions, such as strategic manipulation, fairness sensitivity, or minimal-effort gain, into quantifiable signals over feature-level data. This translation is non-trivial: for instance, estimating *causal impact* requires constructing realistic counterfactuals without violating feature dependencies. *Group shift potential* demands unsupervised cohort modeling and sensitivity to boundary-crossing behavior. *Incentive-based heuristics* rely on domain-specific assumptions that are difficult to generalize. Second, many heuristics must operate under partial observability: ground truth intent labels are unavailable, user motivations are latent, and feature semantics may be ambiguous. Finally, the design space is tightly constrained: signals must be interpretable, model-agnostic, and portable across datasets. These factors together render heuristic design a deeply underconstrained and multi-faceted modeling task that requires balancing abstraction with empirical applicability.

4.1.2 Adaptive Combination of Intent Signals. In real-world data environments, no single heuristic can fully capture the varied scenarios captured in our taxonomy. Some emphasize causal impact, others reflect user effort, social sensitivity, or statistical rarity. Our framework adaptively adjusts the influence of each heuristic depending on the identified manipulation scenario, its downstream impact, and its treatment. Note that depending on the identified intent a certain property, such as fairness, consistency, or accuracy might be the target of an attack. To weight our heuristics, we leverage the following intuitions:

- **Scoring Stability.** Signals are strengthened when they exhibit consistency across repeated or structurally similar anomalies, reinforcing their reliability in pattern-rich environments.
- **Regularized Influence.** Signals that dominate attribution without contributing to meaningful downstream improvements are penalized to avoid overfitting or misleading prioritization.

This lightweight, feedback-driven adjustment allows the scoring mechanism to remain sensitive to context without requiring intent-labeled supervision, enabling graceful adaptation to new domains, shifting distributions, and evolving manipulation tactics.

4.2 Evaluation Plan

Our evaluation is designed to answer two complementary questions: (i) how accurately can we estimate the causal impact of detected feature-level errors on model predictions, and (ii) how effectively can we use heuristic factors to identify the underlying intent behind those errors. Each question requires a different experimental setup for each of the scenarios defined in our taxonomy (Figure 1). To evaluate causal impact estimation, we first rely on injected errors in structured datasets with known clean baselines. This allows us to compare predicted counterfactual shifts with observed outputs after correction, providing a controlled environment for impact validation. Yet, some aspects will be common across all of them. To evaluate intent attribution, we assess how well our aggregated heuristics distinguish between intentional and unintentional manipulations. This involves designing perturbation patterns that mimic user incentive structures and comparing heuristic predictions against injection intent labels.

4.2.1 Datasets. There is no dedicated dataset for intent analysis. We can reuse datasets from fairness and ADS literature [31]. These datasets were selected because they likely contain one or more manipulation scenarios from our taxonomy, such as fairness-driven masking, gain-based falsification, or group-level obfuscation.

- *German Credit Dataset* [13]. This dataset contains information and prediction on the creditworthiness of individuals. Intentional errors in this dataset may include overstated income or suppressed existing credit obligations to increase the chance of loan approval. Such manipulations typically reflect user-level gain-seeking behavior.
- *COMPAS Dataset* [17]. This dataset contains data on recidivism risk and includes sensitive attributes, such as race and sex. Intentional manipulation may involve omission or masking of sensitive attributes to avoid discriminatory outcomes, representing group-level adaptation or fairness-motivated obfuscation.
- *IBM HR Analytics Employee Attrition Dataset* [14] contains employee records including job role, monthly income, performance rating, and years at company. Manipulations may arise here from strategic misreporting of attributes such as qualifications or income, often motivated by career advancement incentives.
- *Adult Income Dataset* [4] contains census data used to predict whether an individual's income exceeds $50K/year. Intentional errors may include inflating education level or misreporting race, or misreporting work hours or marital status.

4.2.2 Assumptions. To operationalize the evaluation, we make the following assumptions: *Error Injection.* We assume that the set of erroneous features $\mathcal{E}(x)$ are available via controlled perturbations, allowing ground truth labeling of errors as either natural (unintentional) or strategic (intentional). *Access to Corrected outputs.* For counterfactual causal analysis, we assume access to the ground truth or corrected version for the output of the ML system.

4.2.3 Baselines. To evaluate the effectiveness of our proposed diagnostic framework, we compare against the following baselines: *Uncertainty-Based Detection*, following ideas from RED [25], we can use the model's prediction confidence or entropy as a proxy for detecting anomalous or unreliable inputs. Inputs with low confidence or high uncertainty are flagged as suspicious (Adversarial or OOD). *Adversarial Attacks and Defenses*, inspired by approaches like CIAI [15], Cartella et al. [6] He et al. [11], we train a classifier to distinguish clean inputs from perturbed (simulated adversarial or noisy) inputs. However, this classifier focuses purely on perturbation detection and does not infer causal impact or user intent. We additionally benchmark against CoEvA2 [8], a search-based adversarial testing framework.

4.3 Limitations and Risks

While our framework introduces a structured, heuristic-driven approach for inferring the intent behind input errors, several limitations and risks remain, both in current capabilities and assumptions.

Dependency on Imperfect Error Detection. We assume that erroneous input values are known in advance and focus exclusively on inferring intent post-hoc. If the error detection stage itself is noisy, then downstream attribution may be meaningless or misleading. For instance, an unflagged manipulated input will go undiagnosed, while a clean input misclassified as erroneous could be wrongly labeled as deceptive. A promising future direction is to treat detection and attribution jointly, simultaneously identifying suspicious values and estimating their likelihood of intent.

Absence of Ground Truth and the Fragility of Proxy-Based Attribution. Intent is unobservable in real-world datasets. In the absence of labeled intent, our framework relies on proxy indicators, i.e., causal impact, rarity, or incentive alignment, to approximate malicious intent. However, these signals may be noisy and context-dependent. This introduces the risk of false positives particularly in high-stakes settings such as hiring or credit approval. Moreover, our evaluation uses simulated strategic attacks, which may not fully reflect real-world incentive structures or behavioral strategies.

Sensitivity to Model Internals and Attribution Noise. Several heuristics depend on feature attribution methods to identify high-impact features. However, such attributions can be unstable, model-specific, or even spurious in tabular domains. A manipulated feature may receive high importance due to model brittleness rather than user strategy. Our framework currently lacks mechanisms to distinguish model-induced attribution artifacts from adversarial salience.

Adversarial Mimicry and Diagnostic Ambiguity. A key challenge arises when adversaries craft errors that closely resemble natural input noise. In *gray-box* scenarios, their mimicry is limited by partial knowledge of typical error patterns. In contrast, *white-box* attackers with full access to the dataset and error distribution can simulate benign behavior more effectively. To address this ambiguity, our framework avoids binary labels and instead produces a continuous intent score. For group-level white-box threats, additional heuristics track cohort-level shifts across decision boundaries. However, when adversarial errors closely align with natural patterns, heuristic discrimination weakens, revealing a fundamental limitation.

5 CONCLUSION

This paper outlines the research plan for distinguishing intentional from unintentional data errors in machine learning pipelines. We propose a diagnostic framework that estimates the likelihood that an input error was intentionally introduced and quantifies its causal influence on model predictions. Our approach integrates feature attribution, causal analysis, and domain-aware factors to support more transparent and accountable decision-making. Future work includes implementing the framework, evaluating it on real-world structured datasets, and benchmarking against baselines [6, 15, 30]. This research aims to advance trustworthiness, interpretability, and robustness in machine learning decision systems.

REFERENCES

[1] Fatemeh Ahmadi, Marc Speckmann, Malte F. Kuhlmann, and Ziawasch Abedjan. 2025. MaTElDa: Multi-Table Error Detection. In *EDBT*. OpenProceedings.org, 364–376.

[2] Yuval Bahat, Michal Irani, and Gregory Shakhnarovich. 2019. Natural and Adversarial Error Detection using Invariance to Image Transformations. *CoRR* abs/1902.00236 (2019).

[3] Vincent Ballet, Xavier Renard, Jonathan Aigrain, Thibault Laugel, Pascal Frossard, and Marcin Detyniecki. 2019. Imperceptible Adversarial Attacks on Tabular Data. *CoRR* abs/1911.03274 (2019).

[4] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[5] Nicholas Carlini and David A. Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 39–57.

[6] Francesco Cartella, Orlando Anunciação, Yuki Funabiki, Daisuke Yamaguchi, Toru Akishita, and Olivier Elshocht. 2021. Adversarial Attacks for Tabular Data: Application to Fraud Detection and Imbalanced Data. In *SafeAI@AAAI (CEUR Workshop Proceedings)*, Vol. 2808. CEUR-WS.org.

[7] Mohsen Dehghankar and Abolfazl Asudeh. 2024. Mining the Minoria: Unknown, Under-represented, and Under-performing Minority Groups. *CoRR* abs/2411.04761 (2024).

[8] Salah Ghamizi, Maxime Cordy, Martin Gubri, Mike Papadakis, Andrey Boytsov, Yves Le Traon, and Anne Goujon. 2020. Search-based adversarial testing and improvement of constrained credit scoring systems. In *ESEC/SIGSOFT FSE*. ACM, 1089–1100.

[9] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *ICLR (Poster)*.

[10] Shubha Guha, Falaah Arif Khan, Julia Stoyanovich, and Sebastian Schelter. 2024. Automated Data Cleaning can Hurt Fairness in Machine Learning-Based Decision Making. *IEEE Trans. Knowl. Data Eng.* 36, 12 (2024), 7368–7379.

[11] Zhipeng He, Chun Ouyang, Laith Alzubaidi, Alistair Barros, and Catarina Moreira. 2025. Investigating imperceptibility of adversarial attacks on tabular data: An empirical analysis. *Intell. Syst. Appl.* 25 (2025), 200461.

[12] Stefan Heindorf, Yan Scholten, Gregor Engels, and Martin Potthast. 2019. Debiasing Vandalism Detection Models at Wikidata. In *GI-Jahrestagung (LNI)*, Vol. P-294. GI, 289–290.

[13] Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

[14] IBM. 2017. IBM HR Analytics Employee Attrition & Performance. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset. Accessed: 2025-05-06.

[15] Anubhooti Jain, Susim Mukul Roy, Kwanit Gupta, Mayank Vatsa, and Richa Singh. 2024. Discerning the Chaos: Detecting Adversarial Perturbations while Disentangling Intentional from Unintentional Noises. In *IJCB*. IEEE, 1–10.

[16] Klim Kireev, Bogdan Kulynych, and Carmela Troncoso. 2023. Adversarial Robustness for Tabular Data through Cost and Utility Awareness. In *NDSS*. The Internet Society.

[17] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. COMPAS Recidivism Dataset. https://github.com/propublica/compas-analysis. ProPublica Investigation.

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards Deep Learning Models Resistant to Adversarial Attacks. *CoRR* abs/1706.06083 (2017).

[19] Mohammad Mahdavi, Ziawasch Abedjan, Raul Castro Fernandez, Samuel Madden, Mourad Ouzzani, Michael Stonebraker, and Nan Tang. 2019. Raha: A Configuration-Free Error Detection System. In *SIGMOD Conference*. ACM, 865–882.

[20] Sedir Mohammed, Felix Naumann, and Hazar Harmouch. 2025. Step-by-Step Data Cleaning Recommendations to Improve ML Prediction Accuracy. In *EDBT*. OpenProceedings.org, 542–554.

[21] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In *CVPR*. IEEE Computer Society, 2574–2582.

[22] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer.* John Wiley & Sons.

[23] Fabio Pierazzi, Feargus Pendlebury, Jacopo Cortellazzi, and Lorenzo Cavallaro. 2020. Intriguing Properties of Adversarial ML Attacks in the Problem Space. In *SP*. IEEE, 1332–1349.

[24] Abdulhakim Ali Qahtan, Ahmed K. Elmagarmid, Raul Castro Fernandez, Mourad Ouzzani, and Nan Tang. 2018. FAHES: A Robust Disguised Missing Values Detector. In *KDD*. ACM, 2100–2109.

[25] Xin Qiu and Risto Miikkulainen. 2022. Detecting Misclassification Errors in Neural Networks with a Gaussian Process Model. In *AAAI*. AAAI Press, 8017–8027.

[26] Theodoros Rekatsinas, Xu Chu, Ihab F. Ilyas, and Christopher Ré. 2017. HoloClean: Holistic Data Repairs with Probabilistic Inference. *Proc. VLDB Endow.* 10, 11 (2017), 1190–1201.

[27] Diego Sáez-Trumper. 2019. Online Disinformation and the Role of Wikipedia. *CoRR* abs/1910.12596 (2019).

[28] Sebastian Schelter, Yuxuan He, Jatin Khilnani, and Julia Stoyanovich. 2020. FairPrep: Promoting Data to a First-Class Citizen in Studies on Fairness-Enhancing Interventions. In *EDBT*. OpenProceedings.org, 395–398.

[29] Sebastian Schelter, Tammo Rukat, and Felix Biessmann. 2021. JENGA - A Framework to Study the Impact of Data Errors on the Predictions of Machine Learning Models. In *EDBT*. OpenProceedings.org, 529–534.

[30] Thibault Simonetto, Salah Ghamizi, Antoine Desjardins, Maxime Cordy, and Yves Le Traon. 2023. Constrained Adaptive Attacks: Realistic Evaluation of Adversarial Examples and Robust Training of Deep Neural Networks for Tabular Data. *CoRR* abs/2311.04503 (2023).

[31] Julia Stoyanovich, Serge Abiteboul, Bill Howe, H. V. Jagadish, and Sebastian Schelter. 2022. Responsible data management. *Commun. ACM* 65, 6 (2022), 64–74. https://doi.org/10.1145/3488717

[32] Mykola Trokhymovych, Muniza Aslam, Ai-Jou Chou, Ricardo Baeza-Yates, and Diego Sáez-Trumper. 2023. Fair Multilingual Vandalism Detection System for Wikipedia. In *KDD*. ACM, 4981–4990.

[33] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan Hines, John Dickerson, and Chirag Shah. 2024. Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. *ACM Comput. Surv.* 56, 12 (2024), 312:1–312:42.