

Article

Detecting and Mitigating Bias in AI Systems: A Methodological Approach to Social Inequality

Elena Grimaccia

Italian National Institute of Statistics, Istat, 00184 Rome, Italy; elena.grimaccia@istat.it

Abstract

Artificial intelligence (AI) systems are increasingly embedded in decisions that affect health, employment, housing, justice, and social policy. Rather than neutral tools, they reflect the data on which they are trained, the objectives they optimize, and the contexts in which they operate, reinforcing and amplifying existing social inequalities. This paper addresses two interrelated questions: how can we detect bias across the AI lifecycle, and what technical and governance strategies can effectively mitigate it? Drawing on critical social science perspectives, empirical case studies from healthcare, housing, policing, and labor markets, and recent advances in fair machine learning (including causal debiasing, synthetic data generation, and fair representation learning), the paper develops a conceptual framework, grounded in sociological theories, spanning data collection, preprocessing, model training, deployment, and post-deployment monitoring. As an original contribution, the paper proposes two complementary analytical frameworks: a Social Harm-Intervention Mapping that recasts mitigation strategies in terms of the types of inequality they address, and an Accountability Matrix that maps agency and accountability gaps across social actors and pipeline stages. The paper argues that these frameworks provide social researchers and policymakers with more actionable guidance than conventional pipeline taxonomies because they foreground the social and institutional dimensions of AI-induced inequality rather than treating bias as a purely technical problem.

Keywords: artificial intelligence; social inequalities; algorithmic bias; fair machine learning; causal models; synthetic data; accountability; ethics; public policy



Academic Editors:

Gabriella Punziano, Felice Addeo,
Angela Delli Paoli and
Giuseppe Masullo

Received: 13 July 2026

Revised: 28 September 2026

Accepted: 28 September 2026

Published: 1 October 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Artificial intelligence has moved rapidly from a niche technical field to a pervasive force shaping diverse domains of human activity. Algorithms now influence decisions about who receives medical care, where policing is intensified, how welfare eligibility is assessed, and which job candidates are shortlisted. In none of these cases is AI neutral: it reflects the data on which it has been trained, the objectives it is designed to optimize, and the institutional contexts in which it is deployed.

This paper is a conceptual study that provides new insights and new ways of looking at existing knowledge. Its aim is to develop a broader conceptual understanding of how AI bias can contribute to the reproduction and exacerbation of social inequalities across the AI lifecycle. To our knowledge, this is the first paper to bring together different bodies of scholarship (the computer-science literature on bias and fairness, the sociological and critical data studies literature on algorithmic harm, the survey-statistics literature on coverage and selection) on the basis of their conceptual relevance to this issue. By connecting these

perspectives, the paper develops new theoretical insights into the mechanisms through which biases arising at different stages of the AI lifecycle may translate into, reproduce, and exacerbate existing social inequalities. The two frameworks proposed in the paper, the Social Harm-Intervention Mapping and the Accountability Matrix, are intended as analytical and hypothesis-generating tools to support further theoretical and empirical investigation of these processes.

Artificial intelligence refers to computational systems capable of performing tasks that typically require human intelligence, such as recognizing patterns, making decisions, and processing natural language. Within AI, machine learning has emerged as the dominant paradigm: algorithms learn statistical regularities from data rather than following explicitly programmed rules. Generative AI models, such as large language models, extend this by producing new content based on learned distributions, and a rapidly growing literature has begun to map the distinctive ethical risks they introduce [1]. These systems are not neutral instruments: they inherit the biases, omissions, and power relations embedded in the data on which they are trained [2]. The result is that AI can simultaneously promise efficiency and equity while systematically producing the opposite for those already marginalized.

The relationship between AI and inequality has attracted growing attention across the social sciences. As Crawford [3] argues, AI is not an abstract technology but a material and social assemblage embedded in specific political economies, whose planetary costs of computation and data extraction fall disproportionately on less powerful communities. Pasquale [4] warns of a “black-box society” in which algorithmic systems govern consequential decisions with no transparency or accountability, while Beer [5] traces how the “data gaze” shapes what can be known and who counts as a legitimate subject of knowledge. Zuboff [6] situates these developments within a broader logic of surveillance capitalism, in which human experience is converted into behavioral data as a raw material for prediction and control.

From a sociological perspective, Grassi’s [7] concept of the “algomorphic society” captures the degree to which algorithms have become structural features of social life, co-constituting identities, opportunities, and power relations. Noble’s [8] analysis of algorithmic oppression shows how search engines and recommendation systems actively reproduce racial and gender hierarchies, while Couldry and Mejias’s [9] critique of data colonialism illuminates how data extraction from marginalized communities mirrors historical patterns of resource appropriation without consent or compensation. Eubanks [10] grounds these arguments empirically, demonstrating how automated systems for welfare eligibility, child welfare, and policing in the United States systematically disadvantage already marginalized groups. These perspectives are not merely theoretical: they help identify which inequalities to look for, why they arise, and what kind of responses are adequate.

At the same time, AI and computational methods offer genuine opportunities for social research: enabling analysis of large scale administrative data, detecting patterns of inequality that would be invisible to traditional survey methods, and supporting audit studies that document discrimination in labor and housing markets. The task is not to reject AI but to subject it to the same critical scrutiny that social scientists apply to any other method or institution.

For social scientists, this raises urgent questions. Does AI risk perpetuating or even amplifying existing social inequalities? And what can social research contribute to detecting and mitigating these harms?

Existing scholarship approaches AI bias from two largely separate directions. Technical surveys catalogue bias types and mitigation methods across the machine-learning lifecycle [2,11,12], while the critical social science literature powerfully diagnoses the harms that biased systems inflict on marginalized groups [3,8,10] without connecting those harms

to stage-specific interventions or to the actors responsible for them. The contribution of this paper lies in bringing these strands together: it offers a full-lifecycle account of bias from an explicitly sociological standpoint, a mapping of mitigation onto types of social harm rather than technical stage, an accountability matrix across social actors and pipeline stages, and the integration of survey-statistical correction methods that are largely absent from the computer-science fairness literature. It is this combination, rather than any single element, that distinguishes the present work from prior surveys and critiques.

On bias mitigation in artificial intelligence, existing research offers firm foundations on which the present work builds. A first, predominantly technical literature organizes mitigation techniques by their location in the machine-learning pipeline [13]. This paper builds on this technical repertoire by adding the social conditions that mitigation is ultimately meant to address: the harms that motivate it, the actors who must enact it, and the sociological mechanisms through which bias becomes inequality. A second, socio-technical strand takes an important step in that direction. In particular, the review by Ahmad et al. [14] couples statistical diagnostics with governance across the AI lifecycle, providing an integrated framework for more equitable AI. Other studies focus on specific domains. Hanna et al. [15], for instance, examine ethical and bias considerations within the clinical and pathology context, categorizing bias into data, development, and interaction sources.

The present paper extends all these contributions by adding an explicitly sociological layer, along three complementary lines. First, it shifts the organizing axis from technical to social harm: rather than sorting interventions by data, algorithm, interface, or governance, it reframes mitigation around the types of harm that biased systems produce as discrimination [8,16,17], exclusion [10,17,18], surveillance and control [6,19,20], self-reinforcing feedback loops [21–23], and epistemic harm [24–26]. Second, it develops an Accountability Matrix that functions as a diagnostic instrument rather than a general discussion of governance, locating agency and accountability gaps across social actors and pipeline stages so that responsibility can be read off directly. Third, it integrates survey-statistical correction methods for under-coverage, measurement error, and informative selection that connect the analysis to established practice in the social sciences. Taken together, these additions are offered as a further step that builds on the previous literature.

This paper develops a methodological framework for detecting and mitigating bias across the AI lifecycle. The first part examines the mechanisms through which AI systems produce and reproduce inequality, tracing these from data collection through to post-deployment feedback. The second part turns to mitigation, reviewing strategies at each stage of the pipeline and drawing on recent technical advances in fair machine learning to show how research can directly serve equity goals. Section 4 proposes two complementary analytical frameworks: a Social Harm-Intervention Mapping and an Accountability Matrix. Section 5 presents illustrative applications in domain-specific challenges.

2. Inequalities in AI Systems

Bias in AI is not a single phenomenon but a family of related problems with distinct origins, mechanisms, and social consequences (Table 1). Bias, unfairness and discrimination are related but not interchangeable [2]. Bias in the statistical sense is a systematic deviation in an estimate or a dataset; it need not be unfair [2,11]. Unfairness is a normative judgment that an outcome wrongly disadvantages a group [27,28]; an unfair outcome is not automatically unlawful discrimination, which is a determination under a particular legal standard [29]. Social inequality, in turn, is broader than any gap in predictive performance [8,10]. The taxonomy that follows gives each form of bias a definition that integrates and builds on existing accounts, such as the lifecycle framework of Suresh & Gutttag [11] and the origin-based review of Mavrogiorgos et al. [12].

Mehrabi et al. [2], in a comprehensive survey of bias and fairness in machine learning, identify more than two dozen distinct forms of bias, spanning data collection, labeling, model design, and deployment. Barocas and Selbst [29] distinguish between disparate treatment, where a system explicitly uses a protected characteristic as an input, and disparate impact, where an apparently neutral system produces discriminatory outcomes. Friedman and Nissenbaum [30], in an early and still influential account, categorized bias in computer systems as preexisting (rooted in social institutions and practices), technical (arising from design choices and computational constraints), and emergent (produced by the interaction of a system with novel contexts of use). Building on these taxonomies, the following types of bias are particularly significant for AI systems operating in social domains.

Algorithmic bias refers to systematic unfairness by design or data that affects outcomes across societal groups; it is the broadest category, encompassing many of the subtypes below [2]. Cognitive bias enters AI through the human choices made in data collection, labeling, and model design: confirmation bias, the tendency to favor data or results that align with prior beliefs, is particularly consequential in research contexts where training datasets are assembled by teams with shared assumptions [31]. Exclusion bias arises when relevant populations or variables are omitted from datasets, producing models that generalize poorly to overlooked groups [16]. Measurement bias reflects inaccurate or skewed data capture: when labels or measures do not accurately reflect the underlying construct, models learn a distorted representation of reality [27]. Out-group homogeneity bias leads to the oversimplification of diverse identities into homogeneous group representations, erasing within group variation. Implicit or prejudice bias reflects societal prejudices encoded in data, reinforcing stereotypes related to gender, race, class, and other characteristics [8]. Recall bias arises from imbalanced or inconsistent labeling, particularly in retrospective data collection. Sampling and selection bias, perhaps the most technically well understood form, occurs when training data are unrepresentative of the population to which the model will be applied, producing systematically worse performance for overlooked groups [32]. Finally, stereotype bias occurs when a model reinforces harmful associations through its learned representations, a pattern extensively documented in word embeddings and generative language models [33].

Table 1. A taxonomy of bias types in AI systems.

Bias Type	Definition/Risk
Algorithmic Bias	Systematic unfairness by design or data; affects outcomes across societal groups [2].
Cognitive Bias	Human reasoning flaws embedded into AI via data handling or modeling choices; includes the biases of developers, labelers, and analysts that shape system design [30,31].
Confirmation Bias	Favoring data or results that align with prior beliefs, limiting the scope of analysis; particularly consequential when training datasets are assembled by teams with shared assumptions [31].
Exclusion Bias	Omission of relevant populations or variables from datasets, causing models to lack generality and perform poorly for underrepresented groups [16].
Measurement Bias	Inaccurate or skewed data capture, where labels or measures do not accurately reflect the underlying construct [27].
Out-group Homogeneity	Oversimplification of diverse identities into homogeneous group representations, treating members of an out-group as more uniform than they are [2].
Prejudice (Implicit) Bias	Societal prejudices reflected in data, reinforcing stereotypes related to gender, race, class, and other characteristics [8].

Table 1. *Cont.*

Bias Type	Definition/Risk
Recall Bias	Retrospective or incompletely documented data systematically under-record certain events or groups, producing a distorted and unstable ground truth that propagates into model error [34].
Sampling/Selection Bias	Unrepresentative samples lead to models that perform poorly on overlooked groups; documented across clinical AI datasets globally [32].
Stereotype Bias	Model reinforces harmful stereotypes through learned associations, documented in word embeddings and generative language models [33].
Evaluation Bias	Arises at the model-evaluation stage when test or benchmark data are unrepresentative of the deployment population, or when aggregate performance metrics conceal systematically worse performance for specific subgroups [2].

These bias types are not mutually exclusive: in practice, they interact with and compound one another. A model trained on administratively collected health data may simultaneously exhibit sampling bias (over-representing patients with regular healthcare access), measurement bias (using cost as a proxy for health need), and stereotype bias (encoding associations between race and risk scores). The following sections trace the pathways through which these biases enter and accumulate across the AI lifecycle.

Bias can also enter at the evaluation stage. A model assessed only on aggregate performance may appear accurate while systematically failing specific subgroups, and benchmark datasets that are themselves unrepresentative can certify a system as fair when it is not; the choice of evaluation metric is therefore itself a potential source of bias [2,11]. This evaluation-stage bias is included in the taxonomy in Table 1 and in the lifecycle shown in Figure 1.

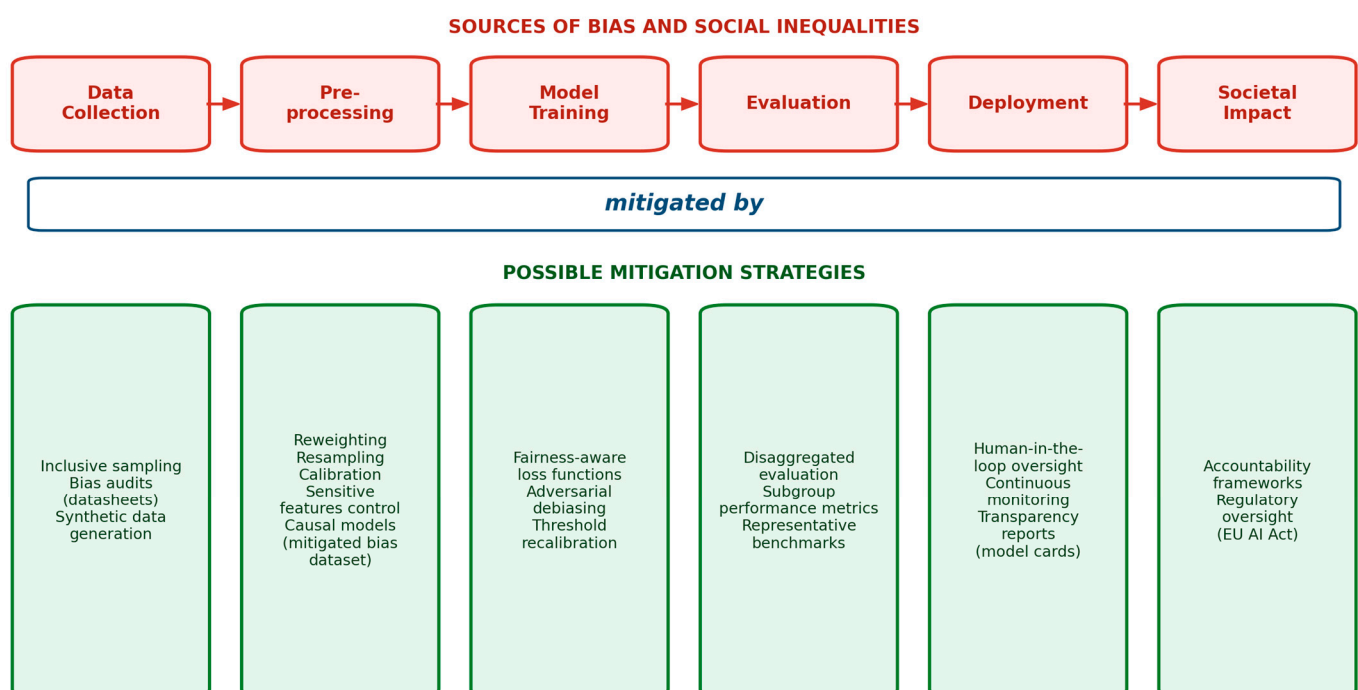


Figure 1. Sources of bias across the AI lifecycle and the corresponding mitigation strategies.

2.1. Data Collection and Representation

Training data are the primary vehicle through which past inequalities enter AI systems. When datasets do not represent all relevant populations, or when they encode

historical patterns of discrimination, models learn those patterns and project them into future decisions.

Facial recognition provides a well-documented illustration. Buolamwini and Gebru [16] found that commercial gender classification systems performed substantially better on lighter skinned men than on darker skinned women, reflecting deep imbalances in the composition of the training data. Buolamwini [35] has continued to document how these disparities are structural: they follow from who is included and who is excluded from the datasets on which systems are trained.

A related problem arises from what Criado-Perez [18] calls the gender data gap: the systematic under representation of women in data collection across health, engineering, and social research. When training data reflect a world measured primarily through a male perspective, models encode that perspective as the default. A recent study on gender bias in large language models applied to long-term care confirms that this dynamic extends to generative AI [36].

Geographic and demographic inequalities in data availability compound these problems. Celi et al. [32] found that the overwhelming majority of clinical AI models relied on datasets from the United States or China, severely limiting their generalizability. This geographical concentration means that clinical AI risks being least accurate precisely in the settings where it is most needed.

Automation bias, i.e., the tendency to over-rely on algorithmic outputs without adequate human correction, is a further mechanism of inequality. Wallace et al. [37] document how automated risk profiling systems for tenant screening in England reproduce existing patterns of socioeconomic disadvantage, presenting discriminatory outputs as objective measures. Cheng et al. [38] similarly find racial disparities in the Allegheny Housing Assessment Tool, where algorithmic assistance in housing allocation systematizes existing inequities.

2.2. Preprocessing and Feature Engineering

Even before model training, design choices in preprocessing can introduce or amplify bias. Proxy variables are a canonical problem: features that appear neutral can serve as surrogates for sensitive attributes. Postcode encodes socioeconomic status and ethnicity; educational institution attended encodes class and race. When such proxies enter a model, the model effectively discriminates on the basis of protected characteristics while appearing not to. Selbst et al. [28] term this the “abstraction trap”: the formalization of fairness requirements without attention to the social context in which data are generated.

Imputation methods present a related risk. When missing data are imputed using majority group patterns, the model is effectively trained to expect minority group members to resemble the majority. Gebru et al. [39] advocate for “datasheets for datasets” as a means of documenting these choices and making them open to scrutiny.

2.3. Model Design and Training

AI models often optimize for aggregate accuracy, concealing substantial heterogeneity in performance across subgroups. Obermeyer et al. [40] provide the clearest demonstration: a healthcare algorithm that predicted costs rather than illness severity systematically underestimated the needs of Black patients, because unequal access to care meant less money was spent on them. Correcting this bias would have increased the proportion of Black patients receiving additional support from 17.7 to 46.5 percent.

Black-box models compound these problems by obscuring the decision making process. Lipton [41] examines the “mythos of model interpretability”, arguing that the opacity of complex models makes it very difficult to identify the source of biased outputs. Hal-

lucination in generative AI raises additional concerns: fabricated outputs may be more consequential for groups already underrepresented in authoritative data sources [33].

2.4. Deployment and Monitoring Risks

Bias also emerges during deployment. Models may be repurposed in ways their developers did not anticipate. Concept drift (the gradual divergence between training data and real world data) can erode performance in ways that fall unequally across groups. Exclusionary design (interfaces that assume high digital literacy or absence of disability) can systematically bar certain populations from AI-assisted services [42].

2.5. Post-Deployment Feedback Loops

Perhaps the most insidious form of AI-induced inequality arises from feedback loops. Ensign et al. [21] show that predictive policing systems trained on historical crime data direct further police attention to already over-policed areas, generating additional recorded incidents that confirm the model's priors. Eubanks [10] documents how automated welfare eligibility systems embed assumptions about poverty that reinforce the conditions that produce it, creating "digital poorhouses". O'Neil [23] argues that these systems function as "weapons of math destruction": they optimize for proxies that encode past inequality and provide no meaningful avenue for appeal.

3. Strategies for Mitigating Bias

Detecting bias is a necessary but not sufficient condition for equity. Effective mitigation requires interventions at each stage of the AI pipeline, combining technical strategies with governance measures. This section surveys those interventions and links them to concrete socio-technical controls and documentation practices that support better governance (Figure 1).

3.1. Technical Strategies

3.1.1. Data Collection

Inclusive sampling is the most direct response to representation bias. A key practical challenge, however, is that sampling strategies can themselves introduce errors when the information used to define strata is inaccurate. Mitani et al. [43] examine this formally in a large multi-center biobank survey enriched for racial and ethnic minorities using electronic health records. They find that under both non-differential and differential misclassification of sampling strata, naive analysis approaches produce biased estimates; appropriately weighted estimators are needed to recover the intended precision gains. This finding is directly relevant to AI training data: enriched sampling of underrepresented groups can backfire if the enrichment strata are themselves measured with error.

Bias audits are the systematic assessments of a dataset's composition and its relationship to outcomes, and they should be conducted before any model training begins. Gebru et al.'s [39] "datasheets for datasets" provides a practical template, making implicit choices visible and open to scrutiny.

Synthetic data generation offers a route to improving minority group representation without compromising individual privacy. Shahul Hameed et al. [44] review synthetic data methods for bias mitigation in healthcare AI, covering generative adversarial networks, variational autoencoders, and agent-based simulation. Their systematic review finds that generated synthetic data can improve accuracy, precision, and fairness, but stresses that effectiveness depends heavily on the quality of the underlying data: poorly generated synthetic data can introduce new artifacts rather than correcting existing biases.

3.1.2. Preprocessing

At the preprocessing stage, reweighting schemes assign differential importance to observations to correct for compositional imbalances [45]; resampling methods alter training data composition by oversampling underrepresented groups [46]. A particularly promising direction is the use of causal models. González-Sendino et al. [47] propose a modified Bayesian network approach that adjusts cause and effect relationships to generate a “mitigated bias dataset”: a synthetic training corpus from which discriminatory pathways have been surgically removed while legitimate predictive relationships are preserved. This approach has three advantages over conventional preprocessing: it is transparent (the causal graph makes explicit which pathways are blocked), proactive (biases are removed before model training), and it preserves sensitive features in the dataset rather than deleting them, thereby avoiding the proxy problem.

3.1.3. Model Training

In-processing methods modify the training objective to incorporate fairness as a constraint. Fairness aware loss functions penalize prediction errors that fall systematically on particular subgroups. Adversarial debiasing methods introduce a secondary network that attempts to predict the sensitive attribute from learned representations; the primary model is trained to minimize its main task loss while also minimizing the adversary’s success [48]. Post-processing methods adjust model outputs after training through threshold recalibration and group-specific corrections [49].

These fairness-aware methods presuppose a definition of fairness, and the choice is neither technical nor neutral. Common criteria (demographic parity, equalized odds, equality of opportunity, calibration, predictive parity, individual fairness, and counterfactual fairness) rest on different assumptions and cannot in general be satisfied simultaneously: formal impossibility results show that calibration and equal error rates are jointly achievable only under equal base rates or perfect prediction [50–52]. Selecting a fairness objective is therefore a normative decision that must be matched to the harm at stake and the decision context: parity-based criteria suit allocation settings where representation is the concern, error-rate criteria suit screening or risk-assessment settings where false positives and false negatives fall unequally on groups, and calibration matters where the score itself is acted upon. Which fairness notion is appropriate is thus a question the Harm-Intervention Mapping helps frame, not one the technical layer can settle on its own.

A recent and technically sophisticated contribution is the FRE framework by Cai and Zhu [53], which decomposes fair representation learning into two progressively decoupled encoding stages. An initial encoder maps features to a latent representation and generates a preliminary label prediction; this prediction is then combined with the original inputs and re-fed into the encoder, producing a final representation that is better calibrated to the task and less dependent on the sensitive attribute. SHAP-based analysis shows that the contribution of the sensitive attribute to target prediction drops by 67 percent on the COMPAS dataset and by over 73 percent on the Adult income dataset after FRE processing. The method’s key limitation is that its counterfactual generation strategy does not yet extend beyond tabular data; and as Hardt et al. [54] note, complete disentanglement of sensitive attributes is not achievable when those attributes are inherently correlated with task-relevant features.

3.1.4. Deployment

At deployment, the most critical safeguards are human-in-the-loop oversight, continuous monitoring, and transparency reporting. The EU AI Act [55] classifies applications in recruitment, credit, and law enforcement as high risk, requiring strict conformity assess-

ments. Explainable AI tools such as SHAP and LIME allow practitioners to understand how model decisions are made for individual cases and communicate findings to non-technical stakeholders. SHAP and LIME provide partial, and sometimes unstable, post hoc explanations of model behavior; they do not establish that a model is fair, that a relationship is causal, or that anyone is accountable, and different explainers (or different runs) can yield divergent attributions. Mitchell et al.'s [42] “model cards” framework provides standardized documentation of model behavior across subgroups, enabling external scrutiny.

A growing ecosystem of open-source toolkits now operationalizes many of the interventions described above. IBM's AI Fairness 360 [56] and Microsoft's Fairlearn [57] bundle standard fairness metrics with pre-, in-, and post-processing mitigation algorithms, and related tools support model auditing and visual inspection¹. Such toolkits lower the practical barrier to fairness testing and make results reproducible, standardizing the technical layer. However, they cannot by themselves determine which harms matter, for whom, or who is accountable. These questions are addressed in the following section.

4. Extension: New Frameworks for Agency and Mitigation Methods

The frameworks proposed in this Section are intended as analytical and hypothesis-generating tools to support further theoretical and empirical investigation of these processes. They are designed to structure inquiry, render visible the relationships between harm production and intervention pathways, and facilitate the formulation of testable propositions for subsequent research.

4.1. A Social Harm-Intervention Mapping

The pipeline-stage taxonomy presented above organizes mitigation strategies by where in the technical workflow they apply. An approach that can be more useful for social scientists and policymakers is to map which population groups, communities, and individuals are most affected by AI-generated biases. We therefore propose reframing the mitigation agenda around five types of social harm that the sociological literature has identified as the primary consequences of biased AI systems: discrimination, exclusion, surveillance and control, feedback loops, and epistemic harm. For each harm, Table 2 identifies the pipeline stages at which it is most acutely generated and the specific interventions best suited to addressing it. The unit of analysis of the Social Harm-Intervention Mapping is a type of social harm; each harm is linked to the pipeline stage(s) at which it is chiefly generated and to the interventions best suited to reducing it. The logic of the mapping is that the appropriate remedy depends on the nature of the problem, not only on its technical location. We organize mitigation around five principal forms of social harm, each grounded in an established body of scholarship: discrimination, where a system produces disparate adverse outcomes for members of a protected group [16]; exclusion, where individuals generate too little data to be “seen” by a system and are thereby denied its benefits [18,19]; surveillance and control, where data-driven monitoring disproportionately targets and disciplines marginalized communities [6,19]; self-reinforcing feedback loops, where biased predictions shape the data subsequently collected and entrench the original disparity [21,22]; and epistemic harm, where systems discredit or displace individuals in their capacity as knowers of their own experience [24,25]. These five categories are intended to capture the principal forms of social harm produced by biased AI systems, each derived from a recognized strand of the literature. They are not offered as an exhaustive typology: the categories may overlap in practice, and further forms of harm may become salient as systems and their social contexts evolve. Our aim here is to analyze mitigation around the harms that recur most consistently across the sociological and computational literature, while leaving the scheme open to extension.

Three observations follow from Table 2. First, the same technical intervention can serve multiple harm-reduction goals: synthetic data generation addresses both exclusion (by improving minority group representation) and discrimination (by providing better training coverage). This convergence suggests that certain investments can yield large positive social returns, particularly in data quality and collection design. Second, some harms, notably surveillance and control and epistemic harm, are not well addressed by in-processing technical fixes; they require governance responses. Third, feedback loops, which are among the most consequential and hardest-to-reverse forms of AI-induced inequality, are addressed primarily, though not exclusively, through post-deployment mechanisms: anticipatory choices at the design and modeling stages can reduce the risk that such dynamics become established, but no amount of preprocessing or fairness-constrained training eliminates a loop driven by the accumulation of AI influenced decisions over time, which is why continued post-deployment monitoring remains essential.

Table 2. Social Harm-Intervention Mapping: Reframing AI bias mitigation around the types of inequality that biased systems produce. Interventions cited are illustrative rather than exhaustive.

Social Harm	Social Manifestation	Primary Pipeline Stage	Key Mitigation Interventions
Discrimination	Differential treatment of individuals on the basis of race, gender, age, disability, or other protected characteristics, producing unequal outcomes in hiring, credit, healthcare, or justice.	Model training; Data collection	Fairness-aware loss functions; adversarial debiasing [48]; fair representation learning [53]; counterfactual fairness [60]; audit studies to detect differential treatment.
Exclusion	Denial of access to services, opportunities, or voice, arising from interfaces that assume high digital literacy, from datasets that erase minority populations, or from models that perform poorly for underrepresented groups.	Data collection; Deployment	Inclusive sampling; active learning to cover edge cases; synthetic data augmentation [44]; accessibility-by-design at deployment; human-in-the-loop review for flagged cases.
Surveillance and control	Asymmetric accumulation of information about individuals and communities, enabling monitoring, profiling, and behavioral control that falls disproportionately on marginalized groups.	Data collection; Deployment	Data minimization and differential privacy; purpose limitation enforced through governance [55]; prohibition of high-risk surveillance uses; community consent and oversight mechanisms.
Feedback loops	Self-reinforcing cycles in which AI-driven decisions shape future data, which in turn confirms and deepens past inequality, as documented in predictive policing [21] and automated welfare systems [10].	Post-deployment; Model training	Continuous monitoring for distributional shift; causal modeling to identify and block discriminatory pathways [47]; regular model retraining with updated data; correction of misclassified strata [43].
Epistemic harm	Suppression or distortion of the knowledge and experiences of marginalized groups, through hallucination that affects underrepresented populations, through training data that encodes a dominant perspective as universal, or through opacity that forecloses contestation.	Data collection; Model design	Datasheets for datasets [39]; explainable AI tools (SHAP, LIME); participatory design involving affected communities; transparency reports [42]; datasheets for models for generative AI.

4.2. Statistical Methods for Mitigating Bias

Beyond the algorithmic fairness techniques discussed in Section 3.1, a parallel body of statistical literature offers critical tools for diagnosing and mitigating bias at the data level. These methods address foundational issues in the data generation process, like under-coverage, measurement error, and non-ignorable selection, that cannot be resolved by in-processing fairness constraints alone.

A core challenge for AI training data is under-coverage, which occurs when certain population units have zero probability of inclusion in the training sample ($p_i = 0$). This problem is pervasive in organically collected datasets (e.g., social media or administrative records) and voluntary opt-in panels, where participation is not controlled by the analyst [61]. Under-coverage can be stochastic or deterministic, with the latter being particularly consequential for fairness because entire subpopulations may be systematically absent from the training data. Survey statisticians have developed correction strategies such as calibration, which reweights sample units using auxiliary population totals to reduce the resulting distortion, and the Missing Information Principle, which provides a framework for inference when data are incomplete [62,63]. In AI contexts, these approaches translate into using external benchmarks to adjust sample weights, ensuring that underrepresented groups receive appropriate influence during model training. Reweighting and calibration, however, can only redistribute weight among groups that are actually observed: where a subpopulation has zero probability of inclusion and is therefore entirely absent from the sample, these adjustments cannot on their own recover it, and valid inference then requires appropriate external information or explicit modeling assumptions.

A related issue is measurement error in the training data. Even when units are included, their observed values may be distorted. In administrative data, this often arises from differing definitions and recording practices across institutions; in opt-in panels, self-reported data can suffer from recall or social desirability biases [64]. Formally, if the true value for a unit in the training set B is y_i , the observed value is $y_i^* = y_i + \varepsilon_i$, where ε_i represents the measurement error [65]. For machine learning, this means that learned associations between features and outcomes may reflect measurement artifacts rather than true relationships. Approaches to address this include error-in-variables models, which explicitly account for the variance introduced by measurement noise, and the use of validation subsamples with gold standard measurements to calibrate the main data [65]. These techniques are essential when training data are assembled from heterogeneous sources, as is increasingly common in large scale AI applications.

Crucially, both under-coverage and measurement error corrections typically assume that selection into the training sample is ignorable after conditioning on observed covariates $\mathbf{x}$ where $P(S = 1 | y, \mathbf{x}) = P(S = 1 | \mathbf{x})$. This assumption, however, is often implausible in practice. Individuals who volunteer for opt-in panels may differ from the broader population in unobservable ways that are also correlated with the outcome of interest (e.g., health status or political engagement), rendering the selection mechanism informative. When ignorability fails, standard adjustments produce biased estimates. Recent methodological innovations offer pathways forward. Pfeiffermann and Sverchkov [66] provide a comprehensive review of approaches for nonprobability samples, emphasizing explicit modeling of the selection mechanism when auxiliary information is insufficient to guarantee ignorability. Chambers [63] extends the Missing Information Principle to accommodate informative selection by leveraging relationships between survey and non-survey data sources, effectively imputing the missing selection probabilities. A particularly promising direction is proposed by Conti and Marella [67], who develop an uncertainty-based framework that quantifies sensitivity to departures from ignorability, allowing analysts to assess how severe the selection bias would have to be to overturn substantive conclusions.

These estimators enter the AI pipeline at two points. During model training, coverage and selection corrections (reweighting, calibration, or adjustment for informative selection) are applied to the training sample before the model is fitted, so that the learned relationships reflect the target population rather than the observed, often non-representative, sample [66,67]. During auditing, the same estimators are applied to the evaluation sample after training: subgroup performance measures such as error and selection rates are re-estimated with survey weights, so that they describe the under-covered groups rather than the audit sample as collected, while sensitivity analysis bounds the uncertainty that remains when coverage is imperfect.

For AI governance, these statistical methods translate into practical requirements. Sensitivity analyses for non-ignorability should be mandated as part of conformity assessments for high-risk systems [55,68]. Datasheets for datasets [39] should document not only what was collected but also known sources of under-coverage and measurement error, alongside the correction strategies applied. Without such documentation, even the most sophisticated fairness-constrained algorithms may simply perpetuate biases encoded in the training data, reinforcing rather than reducing the social harms mapped in Table 2.

4.3. Societal and Governance Dimensions

Technical mitigation strategies are necessary but not sufficient. The Harm-Intervention Mapping shows that some of the most consequential harms, like surveillance, epistemic marginalization, and self-reinforcing feedback loops, require governance responses that no algorithm can substitute for. The Accountability Matrix (Table 3) shows that even the best technical interventions will remain unevenly applied so long as agency is concentrated among actors who face limited accountability. The unit of analysis of the Accountability Matrix is an actor-by-stage pair; each cell records whether an actor typically holds agency (the practical power to act), faces an accountability gap (responsibility without effective power, or power without responsibility), or occupies a partial position in which agency is shared or contested.

Recent work has clarified what accountability means in AI. Building on Bovens's [69] account of accountability as answerability between an agent and a forum, Novelli, Taddeo and Floridi [70] analyze accountability in terms of answerability. They identify three conditions (authority recognition, interrogation, and limitation of power), and seven features (context, range, agent, forum, standards, process, and implications) through four accountability goals (compliance, report, oversight, and enforcement).

The Accountability Matrix proposed here is intended to offer a sociotechnical instrument for identifying, in a concrete system, which agent holds power at each stage and where answerability breaks down: it turns the agent–forum relationship into an applicable tool that maps, across social actors and pipeline stages, where agency lies and where accountability gaps arise. However, the Accountability Matrix is proposed as an analytical approach, a methodological tool for considering simultaneously the role each agent may hold and where the resulting accountability gaps arise [71]. It is a hypothesis-generating device rather than a fixed scheme: agency can differ across sectors, and not every context involves all and only these agents, so the categories it assigns are meant to be adapted to each context, and its rows can be extended or reduced to the agents a given setting makes relevant [70]. Diversity in AI development teams is not merely a matter of equity in the workplace: it is a functional requirement for bias detection, since homogeneous teams share blind spots about whose experiences their systems fail to capture. Human-in-the-loop processes that keep human reviewers involved in consequential decisions provide a mechanism for catching errors that automated monitoring might miss, although their effectiveness is conditional rather than guaranteed: it depends on reviewers' competence

and domain knowledge, their access to the information needed to assess a decision, their authority to override system outputs, and the existence of meaningful review or appeal mechanisms. Community consent (the requirement that those whose data are used and whose lives are affected have meaningful input into system design) is not only an ethical requirement but an epistemic one: affected communities often know better than designers where a system will fail [72].

A second limitation of pipeline-stage taxonomies is that they describe what interventions exist without saying who is responsible for implementing them. This matters enormously in practice. A fairness constraint that an ML engineer could implement in a day may never be applied because the deploying organization did not require it, the regulator did not mandate it, and the affected community had no mechanism to demand it. Conversely, regulatory requirements that are formally in place may be unenforceable if they lack technical specificity or institutional capacity [68].

Table 3 maps agency and accountability across five categories of actor (data producers and collectors, ML engineers, deploying organizations, auditors and regulators, and affected communities) at each stage of the AI pipeline. For each cell, we identify whether the actor currently holds agency (the practical power to act), faces an accountability gap (responsibility without effective power, or power without responsibility), or occupies a partial position in which agency is shared or contested.

Table 3. Accountability Matrix: agency and accountability across social actors and pipeline stages. Green (●) = agency; red (▲) = accountability gap; yellow (⦿) = partial or contested agency. This framework identifies where structural reform is most urgently needed.

Pipeline Stage	Data Producers/Collectors	ML Engineers	Deploying Organizations	Auditors & Regulators	Affected Communities
Data Collection	● Agency: decides what to collect, who to include, how to label.	▲ Gap: rarely involved; receives data as given.	▲ Gap: may commission data but rarely scrutinizes its composition.	▲ Gap: no standard audit requirements at collection stage.	▲ Gap: rarely consulted; data extracted without meaningful consent.
Preprocessing	▲ Gap: upstream; choices already fixed.	● Agency: controls feature selection, imputation, proxy variables.	▲ Gap: typically unaware of preprocessing decisions.	▲ Gap: preprocessing rarely disclosed or audited.	▲ Gap: no visibility into how their data are transformed.
Model Training	▲ Gap: no direct role once data are collected.	● Agency: selects architecture, loss function, fairness constraints.	⦿ Partial: sets objectives; may impose fairness requirements or ignore them.	⦿ Partial: EU AI Act requires conformity assessment for high-risk systems.	▲ Gap: no influence on training process.
Deployment	▲ Gap: no role.	⦿ Partial: may provide technical support; not primarily accountable.	● Agency: full operational control; bears primary accountability for outcomes.	⦿ Partial: can mandate monitoring and transparency reports.	⦿ Partial: subject to system outputs; limited formal redress mechanisms.
Post-Deployment Monitoring	▲ Gap: no role.	⦿ Partial: may be engaged for model updates.	● Agency: should conduct ongoing monitoring; often does not.	● Agency: can require audits, enforce penalties, mandate correction.	▲ Gap: can report harms informally; no structured feedback channel.
Governance & Accountability	▲ Gap: rarely involved in policy-setting.	⦿ Partial: professional codes emerging (ACM, IEEE); weakly enforced.	● Agency: internal governance; commercial incentives may conflict with fairness.	● Agency: key locus of structural accountability; capacity varies widely.	▲ Gap: acknowledged (e.g., EU AI Act) but not granted participatory rights; structurally marginalized.

The matrix reveals a consistent pattern: agency is heavily concentrated in the early technical stages (data collection, preprocessing, model training) among data producers and ML engineers, who typically face few formal accountability requirements. Deploying organizations hold agency at deployment but are often shielded from liability by contractual and regulatory ambiguity. Regulators have formal authority but are structurally reactive. They can mandate audits after harm has occurred, but rarely have the technical capacity to intervene proactively. Affected communities are acknowledged as stakeholders in the wider policy debate around frameworks such as the EU AI Act, but they are not granted participatory rights by it, and they remain structurally marginalized: they are subject to AI driven decisions at every stage while having no formal channel for input at any of them.

This analysis points to two structural reforms that technical mitigation alone cannot achieve. The first is the creation of meaningful participation mechanisms for affected communities at the data collection and system design stages, not as consultation exercises after design decisions have been made, but as substantive input into what data are collected, how sensitive attributes are defined, and what outcomes the system optimizes for. The second is the development of proactive rather than reactive regulation: regulatory frameworks that require conformity assessments before deployment, mandate ongoing bias monitoring as a condition of continued operation, and specify technical standards that are enforceable rather than merely aspirational.

The analysis so far has treated protected characteristics (such as race, gender, or age) as separate and stable attributes; but in practice, they are neither. Social harms frequently arise at their intersections rather than along any single axis [73,74], and they vary across countries, institutions, and historical conditions, so that a characteristic that carries little disadvantage in one setting may be decisive in another. This has direct consequences for mitigation: a strategy that improves an aggregate metric for one group may fail to benefit those at an intersection [75,76]. Taking into account intersectionality therefore raises further issues: subgroup fairness across combinations of attributes rather than one at a time; the small-sample uncertainty that can arise when intersectional subgroups are sparsely represented; and intersectional invisibility, whereby those who belong to several marginalized groups at once are least likely to be measured, represented, or protected. It also complicates the remedies proposed here, since correcting for a harm often requires observing the very sensitive attributes whose collection carries its own risks. Gathering additional sensitive data therefore cannot be presented as an uncomplicated solution, but must itself be weighed as an intervention with costs.

Both frameworks converge on a conclusion that is central to the sociological analysis of AI: technical mitigation strategies are necessary but not sufficient. The Harm-Intervention Mapping shows that some of the most consequential harms (surveillance, epistemic marginalization, self-reinforcing feedback loops) require governance responses that no algorithm can substitute for, while the Accountability Matrix shows that even the best interventions will remain unevenly applied without structural reform of who is accountable for AI outcomes.

5. Domain-Specific Applications

The recognition that automated classification methods can encode bias and require validation emerged shortly after the development of the first machine learning methods. Chiogna, Spiegelhalter, Franklin and Bull [77], in a direct empirical comparison of expert derived and data derived classification trees applied to the diagnosis of congenital heart disease, showed that the two approaches can produce systematically different results, and that the discrepancies are not random noise but carry substantive meaning. The automatically generated tree had learned from a dataset that reflected existing clinical practice,

with all its embedded assumptions, while the expert tree encoded what clinicians believed should determine diagnosis. This gap is, in essence, an early and precise demonstration of measurement bias and cognitive bias operating together in a machine learning context. The concern articulated in 1996 is the direct ancestor of today's fairness-in-machine-learning agenda, and the solution proposed then remains the solution proposed now: not to abandon automated methods, but to subject them to rigorous comparison with human knowledge and to make their assumptions explicit. The three examples below are offered as brief illustrations of a specific form of AI-induced inequality.

5.1. Detecting Potential Exclusion Mechanisms in Environmental Data: Cluster Analysis and Ranking

Algorithmic ranking systems are widely used in policy contexts to order cities, regions, or institutions by composite indices of health, environmental quality, or social development. A recurrent problem is that continuous rankings present differences between units as meaningful even when those differences fall well within the margin of statistical uncertainty, leading to resource allocation decisions based on noise rather than signal. This is itself a form of potential exclusion mechanism: units near the boundary between a “good” and a “bad” rank may be systematically misclassified, and if those boundaries correlate with socioeconomic characteristics, the misclassification falls unequally.

Bottazzi Schenone et al. [78] address this problem by proposing a cluster-based approach to ranking that replaces the continuous ordering with a set of statistically distinguishable groups. Applied to a composite index of air pollution in Italian and European urban areas, the method uses k-means clustering combined with a non-parametric Wilcoxon rank-sum test to identify the maximum number of clusters whose separation is statistically significant. The result is a granular but robust ranking: cities are assigned to pollution level groups, and only differences between groups are treated as real. Within a group, no ordering is imposed.

From a bias mitigation perspective, this approach has an important property: it makes the uncertainty of classification explicit and auditable. A policymaker using the output cannot extract a spurious ranking of cities within the same group, because the method does not produce one. Applied to socioeconomic or health inequality indices, the same logic would prevent the misallocation of targeted interventions to units whose disadvantage is statistically indistinguishable from that of their neighbors.

5.2. Detecting Inequality in Housing Markets

A recent study on Italian student accommodation rental prices [79] offers a methodological illustration of how regression tree analysis such as the CART approach developed by Breiman [80] can serve as a diagnostic tool for detecting bias in statistical modeling. The study compared multiple linear regression, a parametric approach requiring distributional assumptions, with regression trees, a non-parametric method that imposes no such assumptions but provides no statistical inference. While the primary aim was to model determinants of rental prices across 32 Italian university cities, the comparison revealed that the two methods identified different sets of influential predictors: the linear regression selected residents and youth employment rate, while the regression tree identified crime rate, youth employment rate, and average annual wages as most explanatory. This methodological divergence is substantively meaningful because it shows that model specification choices can systematically affect which variables are foregrounded as drivers of inequality. From a bias mitigation perspective, the CART approach provides two valuable detection capabilities: first, it makes explicit the hierarchical structure of predictive relationships, revealing which covariates produce the largest partition of variance and thus warrant closer scrutiny for potential proxy effects; second, it operates without parametric assumptions

about variable distributions, which avoids the distortions that can arise when non-normal variables are transformed or when outliers are absorbed through dummy variables. When employed alongside regression diagnostics, CART thus functions as a sensitivity tool: if the same variables emerge as important across both parametric and non-parametric specifications, confidence in their role increases, whereas divergence signals the need for deeper investigation into measurement error, sampling bias, or proxy variable effects. This diagnostic logic is directly applicable to fairness audits in AI systems.

5.3. Detecting Discrimination Bias in Labor Markets: An Audit Study

Audit studies are among the most rigorous tools available for detecting discrimination bias in real world settings. By submitting fictitious but otherwise identical applications that differ only in a sensitive attribute, researchers can isolate the causal effect of that attribute on a recruiter's behavior, free from the confounding that afflicts observational data [81].

A growing body of audit research has extended this methodology to digital recruitment platforms, where algorithmic systems filter, rank, and recommend candidates before a human recruiter ever sees a profile. Audit studies on LinkedIn and similar platforms have found that gender shapes not only whether a profile is shortlisted but whether it is viewed at all: NLP-based recruitment systems embedded in these platforms can produce unconscious gender bias, detectable through analysis of how male and female profiles in identical roles are scored and recommended differently. Correspondence-audit designs are particularly well suited to detecting gender inequalities, because they isolate the effect of gender from differences in qualifications. A correspondence-audit method of this kind has recently been applied to STEM positions (data analyst, data scientist, software developer) on LinkedIn, submitting matched male and female fictitious profiles to the same positions and analyzing how they are scored, viewed, and contacted [82]. This illustrates how discrimination bias and exclusion bias can be algorithmically embedded and rendered invisible to the actors who appear to be making the choices.

The relevance for bias mitigation is direct. Audit studies provide the empirical baseline against which mitigation techniques such as FRE [53] or adversarial debiasing can be evaluated in deployment, not just on benchmark datasets. Without audit-based measurement, there is no way to know whether a fairness intervention that improves a model's accuracy parity difference in the lab translates into reduced discrimination in the field. As Table 2 indicates, the harms that audits reveal are not fully addressable by in-processing technical fixes alone: they also call for governance responses, including transparency requirements on how algorithmic ranking and recommendation systems are used in recruitment, audit mechanisms that test for differential treatment, and regulatory frameworks that hold automated candidate screening systems to the same anti-discrimination standards as human decision makers.

6. Conclusions

Artificial intelligence is already embedded in consequential decisions across health, housing, justice, labor markets, and social policy. Its potential to amplify existing inequalities is real and well documented. Its potential to support more equitable social analysis is also real, but it is conditional on the quality of data, the care with which models are designed and evaluated, and the governance frameworks within which they operate.

This paper has developed a conceptual synthesis for detecting and mitigating AI bias across the AI lifecycle. On the detection side, bias arises at every stage of the AI lifecycle, from the composition of training data through to the feedback loops that AI-driven decisions create, and understanding it requires both technical analysis and sociological contextualization. On the mitigation side, recent technical advances in causal debiasing, synthetic data generation,

and fair representation learning provide a maturing toolkit; but these tools are only as effective as the governance frameworks within which they are deployed.

The two frameworks proposed here, the Social Harm-Intervention Mapping and the Accountability Matrix, are offered as analytical tools for social scientists and policymakers working in this space. The first extends and enhances the standard engineering question “where in the pipeline does bias arise?” as the sociological question “what kinds of harm does biased AI produce, and what remedies are available?” The second extends the question on “what can be done?” as “who has the power to act, and where is accountability missing?” Together, they translate technically sophisticated literature into frameworks that are actionable for the social researchers and institutional actors who must ultimately make AI equitable. These matrices are offered as conceptual devices to be adapted to, and tested against, specific sectoral and jurisdictional contexts, not as a validated coding of a fixed empirical reality. Their empirical validation is identified as future work.

Some limitations follow from the nature of this contribution. First, the paper is a conceptual synthesis rather than a systematic review: the literature it draws on was selected for theoretical relevance rather than through exhaustive, protocol-driven screening, so its coverage is deliberately selective and other pertinent work will inevitably have been left aside. Second, the two frameworks proposed here (the Social Harm-Intervention Mapping and the Accountability Matrix) are analytical, hypothesis-generating, and they offer actionable guidance but remain to be tested. Third, the frameworks are context-dependent: agency and accountability are not constant across sectors, jurisdictions, procurement arrangements, or organizational forms, and the single status assigned to each actor category is a first-order abstraction meant to be adapted to, and re-examined in, concrete settings. Fourth, the evidence and the governance discussion are geographically concentrated, drawing disproportionately on European and North American cases and on the EU regulatory framework. Future work could focus on extending the results to other, broader experiences and to the empirical validation of both frameworks, including their application across contrasting sectors and jurisdictions.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created in this study. Data sources cited are publicly available through the referenced publications.

Acknowledgments: The author is sincerely grateful to the editors of the Topical Collection “Epistemology, Methodology and Ontology of a Post-Digital Social Inquiry” (Gabriella Punziano, Felice Addeo, Angela Delli Paoli, and Giuseppe Masullo) and the anonymous reviewers for their valuable comments. The article processing charge (APC) was waived through the MDPI Best Paper Award in Digital Social Research received at the VI International Conference ILIS, “Epistemic Shifts in the Post-Digital Society: The Dilemma of Knowledge Production between Generative Models and Quantum Futures,” hosted by the University of Naples Federico II.

Conflicts of Interest: The author declares no conflicts of interest.

Notes

- ¹ Beyond these, the University of Chicago’s Aequitas [58] provides a bias and fairness audit toolkit oriented towards policy and public-sector decision-making, generating group-level bias reports across a range of fairness metrics, while Google’s What-If Tool [59] offers an interactive, largely code-free interface for probing model behavior, testing counterfactuals, and comparing performance across population subgroups. These tools usefully standardize fairness metrics and mitigation algorithms but do not by themselves resolve the social-context and accountability problems the paper foregrounds.

References

- Hagendorff, T. Mapping the ethics of generative AI: A comprehensive scoping review. *arXiv* **2024**, arXiv:2402.08323.
- Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **2021**, *54*, 1–35. [\[CrossRef\]](#)
- Crawford, K. *Atlas of AI*; Yale University Press: New Haven, CT, USA, 2021.
- Pasquale, F. *New Laws of Robotics: Defending Human Expertise in the Age of AI*; Harvard University Press: Cambridge, MA, USA, 2020.
- Beer, D. *The Data Gaze: Capitalism, Power and Perception*; SAGE: Thousand Oaks, CA, USA, 2019.
- Zuboff, S. *The Age of Surveillance Capitalism*; PublicAffairs: New York, NY, USA, 2019.
- Grassi, M. The algomorphic society: Algorithms as social agents. *Sociol. E Politiche Soc.* **2025**, *28*, 45–67.
- Noble, S.U. *Algorithms of Oppression*; NYU Press: New York, NY, USA, 2018.
- Couldry, N.; Meijas, U.A. *The Costs of Connection*; Stanford University Press: Redwood City, CA, USA, 2019.
- Eubanks, V. *Automating Inequality*; St. Martin's Press: New York, NY, USA, 2018.
- Suresh, H.; Guttag, J. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*; ACM: New York, NY, USA, 2021; pp. 1–9.
- Mavrogiorgos, K.; Kiourtis, A.; Mavrogiorgou, A.; Menychtas, A.; Kyriazis, D. Bias in machine learning: A literature review. *Appl. Sci.* **2024**, *14*, 8860. [\[CrossRef\]](#)
- Caton, S.; Haas, C. Fairness in machine learning: A survey. *ACM Comput. Surv.* **2024**, *56*, 166. [\[CrossRef\]](#)
- Ahmad, A.; Vallès, Y.; Idaghdour, Y. Bias in AI systems: Integrating formal and socio-technical approaches. *Front. Big Data* **2026**, *8*, 1686452. [\[CrossRef\]](#)
- Hanna, M.G.; Pantanowitz, L.; Jackson, B.; Palmer, O.; Visweswaran, S.; Pantanowitz, J.; Deebajah, M.; Rashidi, H.H. Ethical and bias considerations in artificial intelligence/machine learning. *Mod. Pathol.* **2025**, *38*, 100686. [\[CrossRef\]](#)
- Buolamwini, J.; Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, New York, NY, USA, 23–24 February 2018; Volume 81, pp. 77–91.
- Tucker, C. *Artificial Intelligence and Algorithmic Exclusion*; The Hamilton Project, Brookings Institution: Washington, DC, USA, 2025.
- Criado-Perez, C. *Invisible Women: Exposing Data Bias in a World Designed for Men*; Vintage: New York, NY, USA, 2019.
- Browne, S. *Dark Matters: On the Surveillance of Blackness*; Duke University Press: Durham, NC, USA, 2015.
- Gandy, O.H. *The Panoptic Sort: A Political Economy of Personal Information*, 2nd ed.; Oxford University Press: Oxford, UK, 2021.
- Ensign, D.; Friedler, S.A.; Neville, S.; Scheidegger, C.; Venkatasubramanian, S. Runaway feedback loops in predictive policing. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, New York, NY, USA, 23–24 February 2018; Volume 81, pp. 160–171.
- Lum, K.; Isaac, W. To predict and serve? *Significance* **2016**, *13*, 14–19. [\[CrossRef\]](#)
- O'Neil, C. *Weapons of Math Destruction*; Crown Publishing Group: Danvers, MA, USA, 2016.
- Fricker, M. *Epistemic Injustice: Power and the Ethics of Knowing*; Oxford University Press: Oxford, UK, 2007.
- Pozzi, G. Automated opioid risk scores: A case for machine learning-induced epistemic injustice in healthcare. *Ethics Inf. Technol.* **2023**, *25*, 3. [\[CrossRef\]](#)
- Symons, J.; Alvarado, R. Epistemic injustice and data science technologies. *Synthese* **2022**, *200*, 87. [\[CrossRef\]](#)
- Jacobs, A.Z.; Wallach, H. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 375–385.
- Selbst, A.D.; Boyd, D.; Friedler, S.A.; Venkatasubramanian, S.; Vertesi, J. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 59–68.
- Barocas, S.; Selbst, A. Big data's disparate impact. *Calif. Law Rev.* **2016**, *104*, 671–732.
- Friedman, B.; Nissenbaum, H. Bias in computer systems. *ACM Trans. Inf. Syst.* **1996**, *14*, 330–347. [\[CrossRef\]](#)
- Kahneman, D. *Thinking, Fast and Slow*; Farrar, Straus and Giroux: New York, NY, USA, 2011.
- Celi, L.A.; Cellini, J.; Charpignon, M.-L.; Dee, E.C.; Dernoncourt, F.; Eber, R.; Mitchell, W.G.; Moukheiber, L.; Schirmer, J.; Situ, J.; et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities: A global review. *PLoS Digit. Health* **2022**, *1*, e0000022. [\[CrossRef\]](#)
- Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 610–623.
- Sylolypavan, A.; Sleeman, D.; Wu, H.; Sim, M. The impact of inconsistent human annotations on AI driven clinical decision making. *npj Digit. Med.* **2023**, *6*, 26. [\[CrossRef\]](#)

35. Buolamwini, J. *Unmasking AI: My Mission to Protect What Is Human in a World of Machines*; Penguin: New York, NY, USA, 2023.
36. Rickman, S. Evaluating gender bias in large language models in long-term care. *BMC Med. Inform. Decis. Mak.* **2025**, *25*, 274. [CrossRef]
37. Wallace, A.; Beer, D.; Burrows, R.; Ciocănel, A.; Cussens, J. Algorithmic tenancies and the ordinal tenant. *Hous. Stud.* **2026**, *41*, 331–351. [CrossRef]
38. Cheng, L.; Drayton, C.; Chouldechova, A.; Vaithianathan, R. Algorithm-assisted decision making and racial disparities in housing: A Study of the Allegheny Housing Assessment Tool. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*; AAAI Press: Washington, DC, USA, 2025.
39. Gebru, T.; Morgenstern, J.; Vecchione, B.; Wortman Vaughan, J.; Wallach, H.; Daumé, H.I.; Crawford, K. Datasheets for Datasets. Communications of the ACM. Acm.Org. December 2021. Available online: <https://cacm.acm.org/research/datasheets-for-datasets/> (accessed on 28 April 2026).
40. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **2019**, *366*, 447–453. [CrossRef]
41. Lipton, Z.C. The mythos of model interpretability. *Queue* **2018**, *16*, 31–57. [CrossRef]
42. Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 220–229.
43. Mitani, A.A.; Mercaldo, N.D.; Haneuse, S.; Schildcrout, J.S. Survey design and analysis considerations when utilizing misclassified sampling strata. *BMC Med. Res. Methodol.* **2023**, *23*, 166.
44. Shahul Hameed, M.A.; Qureshi, A.M.; Kaushik, A. Bias mitigation via synthetic data generation: A review. *Electronics* **2024**, *13*, 3909. [CrossRef]
45. Li, C.; Ding, S.; Zou, N.; Hu, X.; Jiang, X.; Zhang, K. Multi-task learning with dynamic re-weighting to achieve fairness in healthcare predictive modelling. *J. Biomed. Inform.* **2023**, *143*, 104399. [CrossRef]
46. Feng, R.; Calmon, F.P.; Wang, H. Adapting fairness interventions to missing values. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2024; pp. 59388–59409.
47. González-Sendino, R.; Serrano, E.; Bajo, J. Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making. *Future Gener. Comput. Syst.* **2024**, *155*, 384–401. [CrossRef]
48. Zhang, B.H.; Lemoine, B.; Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*; Association for Computing Machinery: New York, NY, USA, 2018; pp. 335–340.
49. Jang, T.; Shi, P.; Wang, X. Group-aware threshold adaptation for fair classification. *Proc. AAAI Conf.* **2022**, *36*, 6988–6995. [CrossRef]
50. Kleinberg, J.; Mullainathan, S.; Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS)*; Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik: Wadern, Germany, 2017; pp. 43:1–43:23. [CrossRef]
51. Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* **2017**, *5*, 153–163. [CrossRef]
52. Castelnovo, A.; Crupi, R.; Greco, G.; Regoli, D.; Penco, I.G.; Cosentini, A.C. A clarification of the nuances in the fairness metrics landscape. *Sci. Rep.* **2022**, *12*, 4209. [CrossRef]
53. Cai, J.; Zhu, F. Mitigate discrimination caused by data bias: Learning fair representations with information reinforcement via recurrent encoding. *Expert Syst. Appl.* **2026**, *297*, 129379. [CrossRef]
54. Hardt, M.; Price, E.; Srebro, N. Equality of opportunity in supervised learning. In *Proceedings of the 29th International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2016.
55. European Union. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*; Official Journal of the European Union: Luxembourg, 2024.
56. Bellamy, R.K.E.; Dey, K.; Hind, M.; Hoffman, S.C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; et al. AI Fairness 360: An extensible toolkit for detecting and understanding unwanted algorithmic bias. *IBM J. Res. Dev.* **2019**, *63*, 4:1–4:20. [CrossRef]
57. Bird, S.; Dudík, M.; Edgar, R.; Horn, B.; Lutz, R.; Milan, V.; Sameki, M.; Wallach, H.M.; Walker, K. *Fairlearn: A Toolkit for Assessing and Improving Fairness in AI*; MSR-TR-2020-32; Microsoft: Redmond, WA, USA, 2020.
58. Saleiro, P.; Kuester, B.; Hinkson, L.; London, J.; Stevens, A.; Anisfeld, A.; Rodolfa, K.T.; Ghani, R. Aequitas: A Bias and Fairness Audit Toolkit. *arXiv* **2018**, arXiv:1811.05577.
59. Wexler, J.; Pushkarna, M.; Bolukbasi, T.; Wattenberg, M.; Viégas, F.; Wilson, J. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Trans. Vis. Comput. Graph.* **2020**, *26*, 56–65. [CrossRef]

60. Kilbertus, N.; Rojas-Carulla, M.; Parascandolo, G.; Hardt, M.; Janzing, D.; Schölkopf, B. Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 656–666.
61. Chen, Y.; Li, P.; Wu, C. Dealing with undercoverage for non-probability survey samples. *Surv. Methodol.* **2023**, *49*, 497–515.
62. Kim, J.-K.; Rao, J.N.K. Data integration for big data analysis in finite population inference. *Int. Stat. Rev.* **2018**, *89*, 382–401.
63. Chambers, R.L. The missing information principle—A paradigm for analysis of messy sample survey data. *Surv. Methodol.* **2023**, *49*, 219–257.
64. Dutwin, D.; Buskirk, T.D. Apples to oranges or gala versus golden delicious? Comparing data quality of nonprobability in samples to low response rate probability samples. *Public Opin. Q.* **2017**, *81*, 213–249.
65. Rao, J.N.K. On making valid inferences by integrating data from surveys and other sources. *Sankhyā Indian J. Stat.* **2021**, *83*, 242–272. [[CrossRef](#)]
66. Pfeffermann, D.; Sverchkov, M. Use of nonprobability samples for official statistics, state of the art. *Surv. Methodol.* **2025**, *51*, 169–196.
67. Conti, P.L.; Marella, D. An uncertainty based approach for dealing with selection bias in non-probability samples. *Int. Stat. Rev.* **2025**, *94*, 264–290. [[CrossRef](#)]
68. OECD. *OECD AI Principles*; OECD Publishing: Paris, France, 2024.
69. Bovens, M. Analysing and assessing accountability: A conceptual framework. *Eur. Law J.* **2007**, *13*, 447–468. [[CrossRef](#)]
70. Novelli, C.; Taddeo, M.; Floridi, L. Accountability in artificial intelligence: What it is and how it works. *AI Soc.* **2024**, *39*, 1871–1882. [[CrossRef](#)]
71. Cobbe, J.; Lee, M.S.A.; Singh, J. Reviewable automated decision-making: A framework for accountable algorithmic systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 598–609. [[CrossRef](#)]
72. UNESCO. *Recommendation on the Ethics of Artificial Intelligence*; UNESCO: Paris, France, 2021.
73. Crenshaw, K. Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *Univ. Chic. Leg. Forum* **1989**, *1989*, 8.
74. Collins, P.H. Intersectionality's definitional dilemmas. *Annu. Rev. Sociol.* **2015**, *41*, 1–20. [[CrossRef](#)]
75. Kearns, M.; Neel, S.; Roth, A.; Wu, Z.S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 2564–2572.
76. Foulds, J.R.; Islam, R.; Keya, K.N.; Pan, S. An intersectional definition of fairness. In *Proceedings of the 2020 IEEE 36th International Conference on Data Engineering (ICDE)*, Dallas, TX, USA, 20–24 April 2020; pp. 1918–1921. [[CrossRef](#)]
77. Chiogna, M.; Spiegelhalter, D.J.; Franklin, R.C.; Bull, K. An empirical comparison of expert-derived and data-derived classification trees. *Stat. Med.* **1996**, *15*, 157–169. [[CrossRef](#)]
78. Bottazzi Schenone, M.; Grimaccia, E.; Vichi, M. A Novel Clustering Method with Maximum Number of Ordered Centroids and Stable Clusters for Optimal Ranking in a Univariate Setting. *Stat. Methods Appl.* **2025**, *34*, 607–637. [[CrossRef](#)]
79. Farnè, M.; Grimaccia, E.; Sonnati, A. Comparing parametric and nonparametric models: An application to university students' rental prices. *Stat. Appl. Ital. J. Appl. Stat.* **2023**, *35*, 235–240. [[CrossRef](#)]
80. Breiman, L.; Friedman, J.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Chapman & Hall/CRC: Boca Raton, FL, USA, 1984.
81. Bertrand, M.; Mullainathan, S. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *Am. Econ. Rev.* **2004**, *94*, 991–1013. [[CrossRef](#)]
82. Grimaccia, E.; Moreschini, I.; D'Amato, M.; Dei Pieri, S.; Ferri, F.; Montepara, D. Correspondence Audit Method for Digital Labour Markets: Gender Disparities in STEM Careers on LinkedIn. *Ital. Econ. J.* **2026**, under review.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.