Energy Research & Social Science 116 (2024) 103653

Contents lists available at ScienceDirect

Energy Research & Social Science

journal homepage: www.elsevier.com/locate/erss

Perspective

Addressing machine learning bias to foster energy justice

Chien-fei Chen [a,*], Rebecca Napolitano [b,c,d], Yuqing Hu [b,c,d], Bandana Kar [e], Bing Yao [f]

[a] Department of Sociology, Anthropology and Criminal Justice, Clemson University, USA
[b] Department of Architectural Engineering, Penn State University, USA
[c] Affiliate of the Center for Socially Responsible Artificial Intelligence, Penn State University, USA
[d] Affiliate of the Institute for Computational and Data Sciences, Penn State University, USA
[e] AAAS STP Fellow at Building Technologies Office, U.S. Department of Energy, USA
[f] Department of Industrial and Systems Engineering, University of Tennessee Knoxville, USA

ARTICLE INFO

Keywords:
Machine learning
AI
Data justice
Energy justice
Data bias
Consequences of data bias

ABSTRACT

Energy justice advocates for the equitable and accessible provision of energy services, mainly focusing on marginalized communities. Adopting machine learning in analyzing energy-related data can unintentionally reinforce social inequalities. This perspective highlights the stages in the machine learning process where biases may emerge, from data collection and model development to deployment. Each phase presents distinct challenges and consequences, ultimately influencing the fairness and accuracy of machine learning models. The ramifications of machine learning bias within the energy sector are profound, encompassing issues such as inequalities, the perpetuation of negative feedback loops, privacy concerns regarding, and economic impacts arising from energy burden and energy poverty. Recognizing and rectifying these biases is imperative for leveraging technology to advance society rather than perpetuating existing injustices. Addressing biases at the intersection of energy justice and machine learning requires a comprehensive approach, acknowledging the interconnectedness of social, economic, and technological factors.

1. Introduction

Machine learning (ML) and Artificial Intelligence (AI) have become integral tools for managing large datasets in energy consumption analysis, optimizing efficiency, and uncovering disparities in energy infrastructure access and distribution. While these approaches offer rapid solutions to researchers, they inadvertently introduce implicit bias. Bias in ML, a subset of AI, refers to the systemic and unjust favoritism and marginalization that can manifest during the conception, implementation, and application of ML systems. These biases may stem from various sources, leading to significant ethical and societal implications. This perspective seeks to explore the nexus between ML bias and energy justice, identify different types of biases, and propose solutions for mitigating ML bias. In leveraging ML and AI for energy-related analyses, it is crucial to recognize and address the unintended biases that may permeate these technologies. Understanding the diverse sources of bias in ML systems is essential for developing strategies that promote fairness and justice. This perspective aims to contribute to the ongoing discourse by shedding light on the intricate relationship between ML bias and energy justice, paving the way for informed discussions and concrete steps towards minimizing biases in ML applications within the energy sector.

1.1. Energy justice and machine learning bias

Energy justice has emerged as a recent focal point for researchers and policymakers, aiming to bolster energy equity [1–3]. It advocates for equal access to affordable and reliable energy supply and services for all, particularly emphasizing the needs of underserved communities with three main components – recognition, distributive, and procedural justice [4,5]. Recognition justice highlights the necessity of acknowledging which groups of society are overlooked or misrepresented. Distributive justice revolves around the ethical principles that should guide resource allocation among society members. Finally, procedural justice emphasizes the integrity of the process leading to decisions [2]. Using ML to analyze energy-related data has the potential to introduce and amplify biases, thereby deepening existing social inequalities. Bias can permeate the entire process, from data collection and model development to model deployment (Fig. 1). Mitigating bias in ML for energy analysis is not merely a technical challenge but imperative for promoting social

* Corresponding author.
E-mail address: chienfc@clemson.edu (C.-f. Chen).

https://doi.org/10.1016/j.erss.2024.103653
Received 30 December 2023; Received in revised form 17 June 2024; Accepted 19 June 2024

justice. Researchers and policymakers must be acutely aware of the potential biases and their wide-ranging implications to tackle these challenges effectively. This paper systematically examines various bias types at different stages of ML applications for energy data analysis and proposes future research directions. Furthermore, this paper underscores the need for proactive measures to counteract bias in ML applications for energy analysis.

2. Overview of machine learning bias

As researchers delve deeper into the complexities of bias in ML for energy systems, it becomes evident that biases can emanate from various sources, spanning the entire data pipeline, algorithmic development, and deployment process, as illustrated in Fig. 1. In the first hurdle, data collection, bias can manifest in three primary stages: data collection, integration, and preprocessing. During the data collection phase, bias may arise from raw data acquisition and the utilization of secondary data. Raw data may inherently contain biases due to measurement inaccuracies, sampling, or selection processes. Similarly, secondary data can introduce biases through sampling and selection procedures. Subsequently, data integration can introduce aggregation bias, while data preprocessing can potentially introduce measurement and aggregation biases. If left unaddressed, these biases can culminate in representation bias, impacting spatial and temporal scales within energy systems.

Moreover, these biases can propagate throughout the ML algorithm development pipeline. The design of ML algorithms' loss and reward functions can also exacerbate overall bias. Careful consideration is needed to ensure these functions do not disproportionately penalize certain groups, promoting unfairness in algorithmic decision-making. Evaluation bias is another concern during model performance assessment, mainly when using metrics like accuracy or imbalanced datasets. During post-development, the ML models' deployment phase introduces its own biases. Domain shift bias occurs when the data distribution of the training model differs from its actual deployment environment.

Additionally, explanation bias can arise from many ML models' inherent ‘black box’ nature, making their decision-making processes opaque. Encouraging interpretable and explainable models and documenting model decisions can help address this concern. In the following section, this paper explains each type of bias and explores their impacts and potential strategies to overcome them. As the diagram in Fig. 2 was constructed qualitatively, the weights of each bias and the width of each flow are represented equally and illustrate a high-level mapping of critical bias points and their potential propagation paths.

2.1. The impacts of data collection bias

This perspective summarizes four types of bias during the data collection that can impact energy justice: sampling, measurement, selection, aggregation, and representation.

2.1.1. Sampling bias

Sampling bias emerges when the collected data fails to comprehensively capture and reflect the characteristics of the entire population, thereby yielding flawed or partial outcomes. For instance, exclusively gathering data on energy consumption in households equipped with smart meters might result in underestimating energy usage within low-income communities lacking such technology. Notably, sampling bias is accentuated in datasets reliant on non-representative sampling, as seen in the U.S. Department of Energy's Building Performance Database [6]. These datasets may inadvertently emphasize specific regions or markets, limiting their representativeness [7]. Sampling bias can also manifest during data preprocessing, particularly when handling missing data. Various methods, such as mean imputation (using the mean in place of missing data), omission or removal of missing data, and experience-based assumptions about the randomness of missing data (filling missing data based on past knowledge or experience), are employed to address this issue. However, these methods can inadvertently introduce bias. For instance, mean imputation replaces missing values with the variable's mean from available data, potentially introducing bias if the missing data are neither randomly distributed nor accurately represented. For example, higher-income groups may be more hesitant to disclose their income, leading to a sampling bias. This can result in the misallocation of resources and misalignement of policies intended to improve energy affordability and accessibility. Such misallocation may lead to inefficient use of subsidies and support programs, failing to reach the households most in need and thus undermining the objectives of energy justice. Likewise, sampling bias can adversely impact equitable planning of clean transportation technologies. When city planners utilize surveys to gauge the need for electric vehicle (EV) charging infrastructures, they often unintentionally target EV owners and higher-income individuals who have more access to the technology. This tendency can skew the survey findings, potentially failing to capture the broader communities' needs, especially those of lower-income households. Research has demonstrated that EV charging stations are

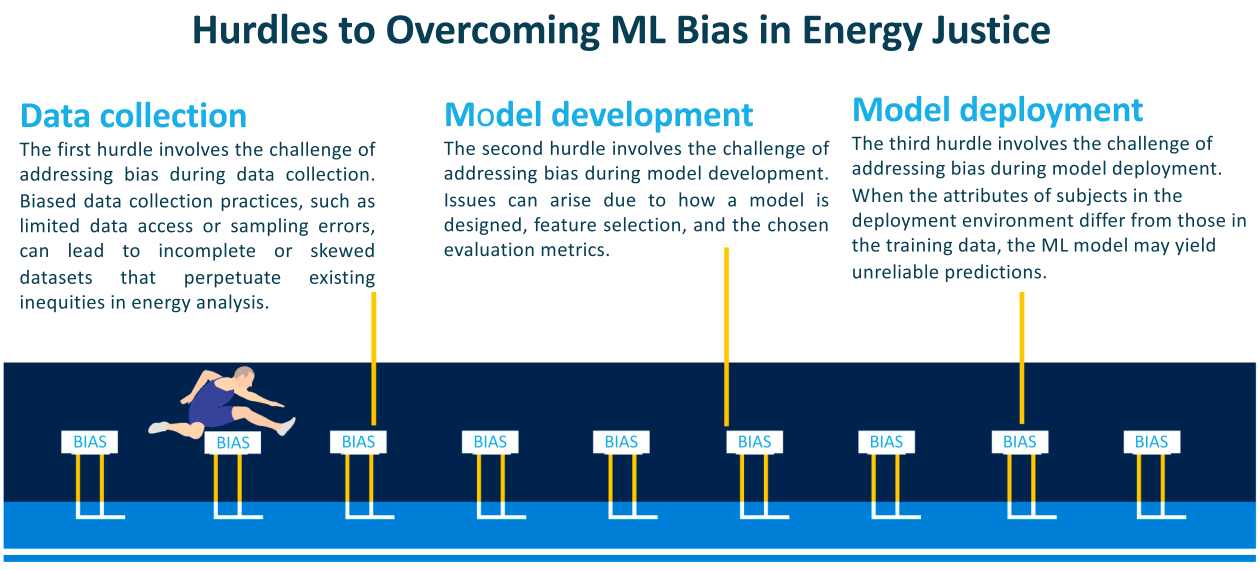

Fig. 1. Three main hurdles to overcoming ML bias relating to energy justice: Data collection, model development, and model deployment.

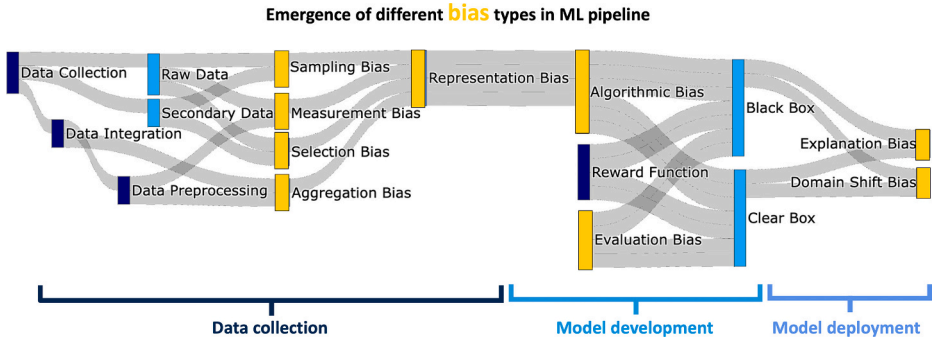

Fig. 2. Connecting hurdles, their origins, and their impacts throughout the ML cycle.

markedly underrepresented in low-income, Black-identifying, and underinvested neighborhoods, thereby limiting access to EV technology for these communities [8]. Hence, such sampling bias might lead city planners to overestimate the demand for EV charging infrastructures for wealthier neighorhoods based on the responses of higher-income participants. This could result in an unequal allocation of resources towards EV infrastructures, diverting funds from vital public transportation projects that are crucial for lower-income residents who depend heavily on them. Inadequate investment in public transportation can severely limit their mobility, making it harder to access jobs, educational opportunities, and essential services. Therefore, sampling bias not only hinders the generalization of results to a broader population but also underscores the necessity to address and mitigate bias at the data collection and processing stages.

2.1.2. *Selection bias*

Selection bias is a broader concept encompassing various biases arising from how subjects are chosen or included in a study. It includes not only sampling bias but also other biases that may occur during the assignment of subjects to groups or the handling of data. For example, in the U.S. national survey data like the Residential Energy Consumption Survey (RECS) [9], intentional exclusion of vacant homes and specific quarters can introduce selection bias. Selection bias also arises when only participants in an energy efficiency program are considered, excluding those who are either unaware of these energy efficiency programs or cannot afford them without financial assistance due to high initial costs or lack of access to financing options. As a result, policies and incentives designed to encourage energy efficiency technology adoptions may be tailored to the needs and behaviors of early adopters, typically from more financially secure groups, rather than addressing the barriers faced by lower-income households to adopt the technologies.

Selection bias can manifest in ML applications when training data are biased towards specific demographic characteristics within a community. Subsequently, this bias introduces social, ethical, and privacy biases, perpetuating stereotypes and discrimination. ML algorithms, such as sentiment analysis and decision trees, are particularly susceptible to selection bias, leading to biased outcomes and inaccurate classifications. Even support vector machines, if trained on data biased towards affluent neighborhoods, can perpetuate biases and contribute to inadequate energy resource allocation for low-income or diverse communities, reinforcing existing inequalities.

2.1.3. *Measurement bias*

Measurement bias arises when collected data is systematically inaccurate or incomplete, distorting analyses. This bias can originate from restricted data access or inadequate instrumentation in energy analysis. An illustrative sample is measurement validity, which gauges the extent to which an assessment tool, such as a questionnaire, accurately measures its intended construct. Without validity, results obtained from the measurement or tool may fail to faithfully represent the underlying construct of interest. Typically, validity is established through empirical evidence and theoretical reasoning [10]. Assumptions grounded in experience can contribute to measurement bias. Considering studies using the American Time Use Survey (ATUS) dataset, which lacks finer location data (i.e., only regional data without census tract or zip code), the absence of finer geographic identifiers poses a significant barrier to understanding energy consumption patterns. To compensate for the lack of fine geolocation, researchers must make assumptions about the locations of household activities for building energy consumption simulations. For instance, many building energy consumption models might assume that work-related energy use predominantly occurs outside the home. However, due to the COVID-19 pandemic, the increase in work-from-home arrangements has started altering this pattern. Suppose energy policy planners use outdated models, such as assuming energy consumption occurs in office buildings, without considering this shift, they may implement rolling blackout plans or energy distribution strategies that inaccurately predict energy demand peaks and valleys. Such measurement bias can disproportionately affect residential areas with higher concentrations of remote workers. Several ML algorithms are susceptible to measurement bias. For example, decision tree algorithms such as Classification and Regression Trees (CART) and Random Forests utilize data to create decision rules for prediction, classification, and pattern recognition. If the data used to build these decision trees is biased or incomplete, it could result in skewed rules, potentially leading to misclassification or misprediction for specific populations or neighborhoods. Similarly, linear regression models that rely on historical energy consumption data to predict the general population's future energy use or patterns may underestimate the behavioral patterns and needs of underrepresented communities. Therefore, addressing measurement bias is crucial for ensuring the accuracy and fairness of analyses and predictions.

2.1.4. *Aggregation bias*

This bias emerges when data is consolidated or averaged at a higher level, such as a group or population, resulting in unintentional assumptions or inferences about individuals within that group [11]. Aggregation bias can also manifest when researchers amalgamate variables to have an aggregated score or index. The consequences of aggregation bias are particularly profound in the context of energy justice, as it engenders inaccurate or biased conclusions regarding the unique energy needs and usage patterns of distinct subgroups. For instance, assessing renewable energy adoption rates across affluent and low-income areas may obscure the specific impacts on low-income communities, perpetuating existing inequalities. This aggregation bias, in turn, can lead to misallocation of resources, formulation of misguided policies, and further marginalization of affected communities. ML models focus on evaluating energy affordability, relying solely on aggregated averages, risk overlooking instances of extreme energy poverty, or underestimating the financial burdens faced by low-income or rural households. The issue of aggregation bias is exacerbated during data integration, as illustrated using composite indexes like the Social

Vulnerability Index from the U.S. Center for Disease Control and Prevention (CDC) [12], which combines multiple variables. Such an index may neglect the distinct impact of individual variables. When data from different subgroups are aggregated, cluster analysis may yield biased results. For instance, clustering energy usage patterns or renewable energy adoption without segregating data from diverse subgroups can lead to inaccurate representations of their specific challenges.

Modifiable Areal Unit Problem (MAUP) is a well-known statistical bias resulting from aggregating data across different geographic boundaries, eventually leading to various patterns and measurements [13]. For instance, energy burden data from the Low-income Energy Affordability Tool (LEAD) are available at PUMA (Public Use Microdata Areas - the lowest U.S. Census spatial unit) scale, but the socioeconomic datasets from the Census tend to be available at census blocks or block groups or other boundaries. Given that the boundaries of these two spatial units do not reconcile, they must be aggregated to a common boundary, such as census tracts, counties, or utility service areas. Thus, a study using census tracts vs. counties will have different results due to MAUP [14]. Likewise, power outage data are available from utilities at utility service area boundaries, which differ from county boundaries and may overlap different zip codes. Such data, therefore, must be disaggregated/aggregated to county boundaries or other census boundaries to undertake power restoration and resource planning activities that may marginalize specific communities and exacerbate energy justice [15]. Analogous to MAUP is the Modifiable Temporal Unit Problem (MTUP) that results from the aggregation of time-varying datasets, such as power outage data that can be aggregated hourly or daily depending upon the need [16]. Because of MAUP and MTUP, ML/AI algorithms can produce biased results when data from different subgroups are aggregated at varying spatial and temporal scales, leading to inaccurate representation of specific subgroups' needs and challenges, influencing resource allocation.

2.1.5. Representation bias

Representation bias comprises ***two*** key facets: spatial and temporal bias. Spatial representation bias arises when national datasets are employed locally, introducing disparities in representation due to MAUP. For example, researchers often use the ATUS dataset to analyze occupant energy use behavior patterns in residential building energy use simulation [17,18] or EV charging pattern analysis [19]. Clustering (e.g., K-means) and Markov-Chain are used to identify and simulate these behavior patterns. These pattern identification and simulations are conducted based on the national data. However, challenges emerge when applying these models directly to a specific area (e.g., a state or a city) due to representation issues during the spatial downscaling process, where occupant profiles vary significantly across different regions. Temporal-based representation bias can stem from infrequent data collection or using data collected at one temporal scale to analyze patterns at a different scale. This approach can hinder the accurate capture of short-term changes [20] and lead to MTUP, thereby profoundly impacting energy justice. For example, low-income households often lack access to smart meters, which track detailed, real-time energy consumption. Consequently, their energy usage data lacks the granularity necessary to capture short-term energy use peaks accompanying extreme weather, e.g., heat waves or extreme cold. Without detailed data, energy assistance programs might distribute resources based on average monthly usage, missing the urgent needs of these households in critical times. To address temporal bias, social-media data (e.g., Twitter (aka X)) or other real-time datasets (e.g., SafeGraph) are often incorporated to analyze short-term changes, especially during emergencies, such as power outages [21,22]. However, it is essential to note that such datasets are accessible only to a certain percentage of the population using the respective apps or devices and willing to share their locations and associated information. These biases adversely affect decision tree algorithms, neural networks, and clustering algorithms, leading to inaccurate predictions, unfair classifications, biased outputs, and insufficient resource allocation.

2.2. The impacts of model development and deployment

While datasets are the basis for bias, it is crucial to recognize that ML models can also introduce biases even when the training data is inherently unbiased [24]. Algorithmic bias, called model bias, emerges in model development due to model design, feature selection, objective function definitions, and the chosen evaluation metrics. For example, supervised machine learning algorithms, operating on labeled data where each input example is associated with a corresponding target output, are often trained by minimizing the discrepancy between the predicted and ground-truth label information. An objective function designed to maximize overall predictive performance can inadvertently introduce bias, favoring individuals from majority groups who are more prevalent in the dataset, while neglecting those from racially or socioeconomically underrepresented groups. Similarly, a poorly designed reward function in reinforcement learning (RL) will likely amplify energy disparities. For example, when developing energy efficiency retrofit plans in residential buildings to reduce energy consumption and greenhouse gas emissions, if the reward function in RL is biased towards maximizing cost savings without considering the feasibility or affordability of retrofit measures for low-income households, the learned policy may prioritize upgrades that are inaccessible or unaffordable for marginalized communities. This bias will widen the energy efficiency gap between affluent and disadvantaged neighborhoods, exacerbating energy and environmental injustices.

Moreover, feature selection also plays an essential role in designing unbiased models. For example, when identifying areas for infrastructure investment (e.g., renewable energy installations), if the clustering algorithm, a typical unsupervised learning approach, is biased towards prioritizing features that correlate with higher property values or socioeconomic status, the identified areas for investment may disproportionately benefit affluent neighborhoods. This bias could exacerbate unequal spatial access to essential infrastructures, further perpetuating energy access and resilience disparities. As such, a poorly designed model with inappropriate features or objective/reward functions will result in biased predictive outcomes, eroding confidence in ML/AI-supported decision-making processes among policymakers [25].

Evaluation metrics, which furnish quantitative measures to gauge the performance of ML models, can introduce an additional layer of bias into the model development phase if not tailored to the specific data or task requirements. Models learn and are optimized using training data, and their performance is assessed based on specific chosen metrics. Bias emerges when the selected metrics are not aligned with how the model is developed. A concrete example is when a sole accuracy measure is employed to evaluate a supervised ML model's performance distinguishing between traditional and renewable energy sources. This evaluation approach will introduce significant bias when dealing with imbalanced datasets [26]. For instance, even if a dataset is representative of a real-world scenario, if it comprises 95 % traditional samples and only 5 % renewable samples, the overall prediction accuracy can be as high as 95 % if the model simply predicts all samples as conventional energy sources. However, this model is fundamentally flawed because the prediction accuracy for the renewable samples would be zero. Relying solely on overall accuracy as a metric is unreliable for assessing the performance of supervised ML models in the context of imbalanced datasets [27].

Similarly, evaluation metrics can introduce bias to unsupervised learning algorithms. For example, when analyzing energy consumption patterns across communities, evaluation bias can occur if the performance of the pattern recognition algorithms is assessed solely based on metrics such as cluster purity. Suppose those metrics prioritize the homogeneity of clusters without considering the diversity of energy consumption behaviors across different socioeconomic groups. In that case, the resulting clusters may not accurately reflect the needs and

challenges faced by marginalized communities. Another example is in RL algorithms. When designing optimal policy to allocate resources for energy assistance programs (e.g., subsidies for utility bills, weatherization initiatives), if the performance of the RL is assessed solely based on metrics such as program efficiency without considering the effectiveness or equity of resource allocation, the resulting assistance programs may underserve or exclude racial or socioeconomic minorities. As a best practice, a comprehensive evaluation strategy utilizing various metrics that incorporate diversity and equity considerations across different racial and socioeconomic groups is essential to mitigate potential bias in algorithmic evaluation for energy-related analyses.

Bias can also result during model deployment. Specifically, domain shift and explanation bias are two types of biases that can impact justice in the model deployment phases. A foundational assumption in developing and applying ML models is that the training and testing data stem from independent and identically distributed sources [28]. When the attributes of subjects in the deployment environment differ from those in the training data, the ML model may yield unreliable predictions. For instance, if a model is trained on energy usage data from a high-income community, it may produce inaccurate or conflicting predictions when applied to underserved communities, exacerbating existing energy disparities. Deployment bias may arise from the opaque nature of many sophisticated ML models, such as deep learning models. The limited transparency makes it challenging to interpret the model's predictions, especially for policymakers without a background in data science. Without clear explanations regarding how data inputs influence predictions, deploying AI/ML models can introduce unintended harm. Therefore, explainable AI techniques, domain adaption, and transfer learning methods are critical to promote the development of more responsible and socially beneficial AI/ML systems.

It is important to note that biases in ML are centered on ethical, social, and fairness considerations inherent in the development and deployment of ML technologies. Alongside these biases, numerous other common issues exist, such as reproduction, replication, and data leakage [29–31]. These must be diligently addressed to uphold the validity and reliability of ML algotihms' findings, further advancing and facilitating energy justice.

Table 1 presents the overview of critical concepts in energy justice and their associations with types of bias, with some examples. One thing to highlight is that the three energy justice types are inherently interrelated and can influence one another. The biases listed under each category represent their primary impact areas within the energy justice framework. However, it is crucial to understand that these biases can and often do cross over, affecting multiple aspects of energy justice simultaneously.

Table 1
Summary of types of energy justice and associated bias with examples.

Type of energy justice	Definition	Potential Data and ML Bias	Examples
Recognition	Highlights the necessity of acknowledging which groups of society are overlooked or misrepresented. It promotes equity by addressing both historical and current disparities.	Sampling Selection Representation Aggregation	Omitting to collect samples from underrepresented or marginalized groups compromises the integrity of a general population study
Distributive	It revolves around the ethical principles that should guide resource allocation among society members. Its goal is to ensure a fair allocation of benefits or burdens among various users within a community or society.	Selection Measurement Algorithmic Evaluation	Collecting energy consumption data exclusively from households equipped with smart meters may lead to underestimating energy usage in households without this technology. Omitting key variables and concepts that could estimate the outcomes leads to an inaccurate assessment of unfair distribution.
Procedural	Emphasizes the integrity of the process leading to decisions. It focuses on the fairness of the methods, mechanisms, and procedures used in decision-making rather than the equity of the outcomes (which falls under the focus of distributive justice).	Measurement Aggregation Explanation Domain-shift	Collecting information exclusively from decision-makers while neglecting non-decision-makers' perspectives skews the understanding of the situation. Failing to gather information on critical variables essential for estimating the fairness of the decision-making process results in an incomplete analysis.

3. Consequences of ML Bias

Addressing the consequences of bias in ML for energy analysis is imperative. Failure to recognize and rectify these biases perpetuates systemic inequalities, widening the gaps between privileged and marginalized communities. This paper identifies three consequences of ML bias in the energy sector (see Figs. 3 and 4).

3.1. Discrimination and inequality

Biased ML systems can perpetuate discrimination based on race/ethnicity, gender, and age, further entrenching societal disparities. For example, systems that inadvertently discriminate based on race/ethnicity in determining credit eligibility for energy efficiency loans or clean energy financing can lead to distributive injustice, denying marginalized groups equal access to these resources. Similarly, if ML models automating utility customer service interactions unintentionally exhibit racial biases, it represents a procedural injustice where specific communities do not receive fair treatment.

3.2. Negative feedback loops on underserved communities

A negative feedback loop refers to a cyclical process in which the consequences or effects of an initial factor tend to amplify and reinforce that factor over time, creating a self-perpetuating cycle that intensifies the original issue. In the context of energy justice for economically disadvantaged neighborhoods, residents often face higher energy burdens (the percentage of household income spent on energy bills). This high energy burden can force families to make difficult trade-offs between paying for energy and other essential needs like food, healthcare, or education. Consequently, they may adopt energy-saving practices that compromise their health and well-being, such as underheating or undercooling their homes [32]. These substandard living conditions can lead to adverse health effects, missed school or workdays, and reduced productivity, further straining the household's finances. With limited resources, investing in energy-efficient upgrades or transitioning to clean energy sources becomes increasingly challenging, perpetuating their reliance on older, inefficient, and potentially polluting energy systems. This vicious cycle of high energy burdens compromised living standards, and lack of access to clean energy resources can entrench these communities in a state of energy poverty and environmental injustice, making it difficult to break free from this negative feedback loop without external interventions or assistance.

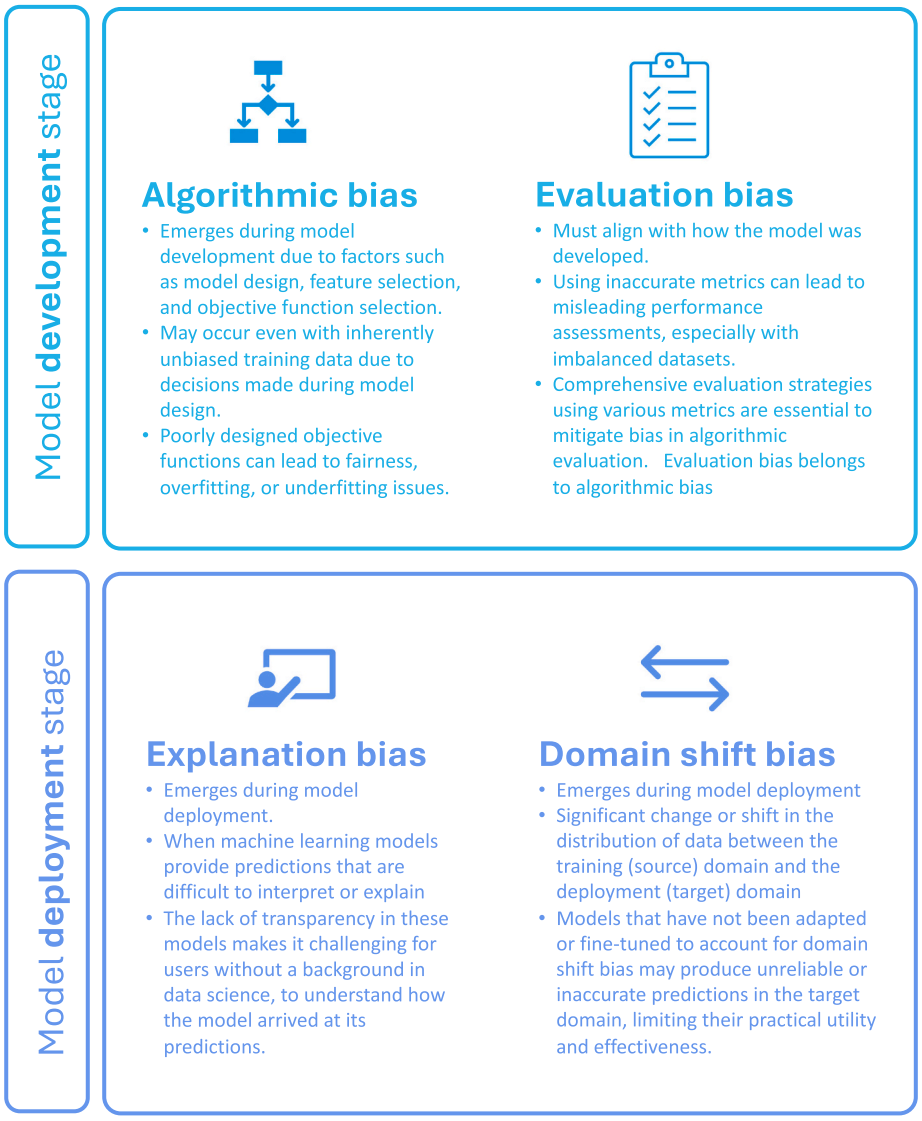

Fig. 3. Biases arising during model development and model deployment stages.

3.3. Inequitable access to advanced technology and economic disparities

Injecting bias into any part of the technology adoption cycle–from assessing eligibility to marketing/education to pricing–can entrench discriminatory barriers that block equitable access to sustainable energy solutions and their associated economic opportunities. Biased ML and customer segmentation models may also systematically deprioritize marketing and outreach efforts in low-income areas, rural communities, or minority neighborhoods. As a result, residents may lack awareness about energy audit programs, rebates, or favorable financing options that could reduce household energy burdens and enable sustainable technology adoption. With the rise of rooftop solar, energy consumers are evolving into prosumers who not only consume but also produce their own electricity through solar energy. This shift has fostered a strong sense of community among energy users. However, during extreme weather conditions such as intense heat or severe cold, when energy demand spikes, these prosumers may encounter difficulties. If their solar production fails to meet the heightened demand, they could face elevated energy costs or even suffer power outages. In these critical times, the ability to supplement their solar supply with energy from the grid or alternative sources becomes crucial.

4. Looking ahead: Strategies for reducing bias

Anticipating bias-free data is unrealistic due to the diverse techniques employed in collecting socioeconomic, weather, physical, built environment, and energy data across varying social, spatial, and temporal scales from diverse sources in different formats. The imperative is minimizing bias to ensure fairness, equity, and accuracy in models. To effectively mitigate bias in ML systems, we propose a comprehensive set of strategies (see Fig. 5) aligned with current best practice guidelines in the ML and AI spheres, such as Fairness, Accountability, Transparency, and Ethics (FATE) [38]. Several toolkits and libraries aligned with FATE principles have emerged to aid in responsible AI development, such as IBM's AI Fairness 360, Microsoft's Fairlearn [39], and Google's ML Fairness Gym [40]. Our proposed strategies build upon these existing FATE-based approaches while tailoring them to the unique challenges at the intersection of machine learning and energy justice:

First, this paper underscores the significance of diverse and inclusive data collection, a core tenet of ethical and fair AI development under the FATE framework. Building a training dataset that encompasses a broad spectrum of demographics and geographic regions is paramount for mitigating biases that can perpetuate injustices, particularly those affecting marginalized groups.

This ethical data collection approach ensures the dataset

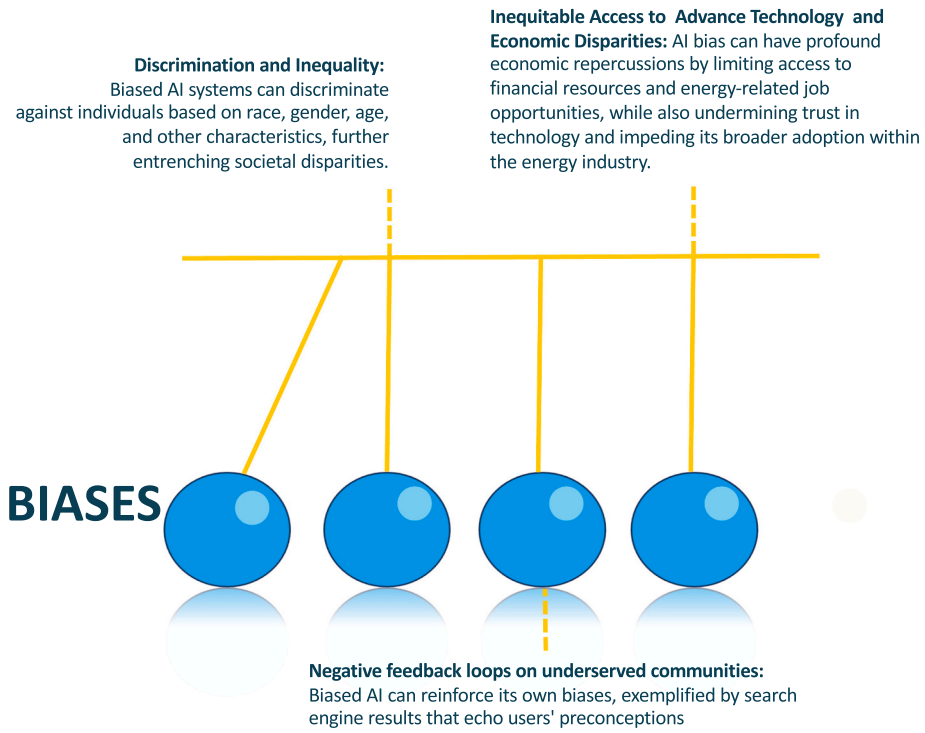

Fig. 4. Consequences of AI Bias in Energy Justice.

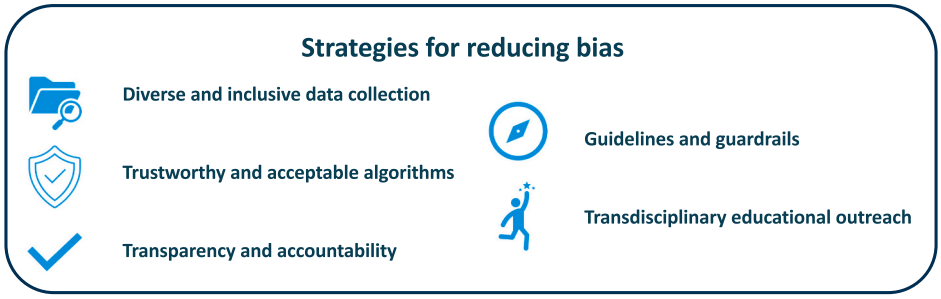

Fig. 5. Strategies for reducing bias.

authentically reflects the characteristics of all target population segments. It advocates for the inclusive representation and participation of communities frequently underrepresented in energy data, upholding principles of fairness and non-discrimination. Moreover, the intentional inclusion of socioeconomic indicators such as income levels, educational attainment, racial/ethnic composition, and access to energy resources [34] contributes to a more equitable and holistic understanding of energy consumption patterns across diverse populations. Capturing this nuanced data is essential for developing fair and ethical ML models that properly account for disparities in energy needs, burdens, and constraints disadvantaged groups face.

Second, this paper advocates for **trustworthy and acceptable algorithms**. Continuous assessment of model performance throughout its lifecycle is crucial to enhance fairness and acceptability among various user groups. Employing debiasing techniques is essential in our pursuit of equitable outcomes. Preprocessing bias mitigation algorithms like those from IBM's AI Fairness 360 toolkit [41] should be used to rectify inadvertent biases introduced during data collection. Bias-aware algorithms, equipped with mechanisms to handle imbalanced data and fine-tune sensitivity to specific groups, also play a crucial role. Post-processing algorithms should recalibrate and transform the model to address concerns about fairness. Evaluation techniques such as Disparate Impact Analysis, Equal Opportunity Analysis, and Confusion Matrix Disparities assess the model's performance across demographic groups. Such practice aligns with fairness evaluation defined in the FATE framework [38]. Leveraging this framework within the context of energy justice is crucial for detecting and addressing energy injustices that could arise from biased model predictions or skewed resource allocations. A point to note is that the debiasing techniques identified here as part of the FATE framework are tailored to minimize bias resulting from algorithms and underlying datasets. Given the data-driven nature of AI, these techniques are designed to enhance the fairness and trustworthiness of the resulting algorithms and outputs. However, they do not capture and minimize data justice issues, a structural challenge requiring human interaction and policies to ensure every household/customer is represented. In other words, debiasing techniques are central to achieving procedural justice, but they may fail only to share and benefit from the data.

Third, this paper argues **transparency and accountability** are central to increasing ML use in the energy sector. These principles align with the FATE framework that has emerged in the trustworthy AI community [38]. The FATE model provides guidelines and tools for developing AI/ML systems that are fair, accountable to stakeholders, transparent in their operation, and uphold ethical principles like non-discrimination and privacy protection. Adopting a FATE-aligned approach can help energy organizations proactively bake in justice

considerations throughout the ML lifecycle. For example, feature importance analysis can help explain model behavior and interpret its performance, aligning with transparency principles from the FATE framework. Continuous monitoring and human involvement throughout the ML model's lifecycle are necessary for achieving transparency and accountability. Domain experts should actively participate in model development and deployment to identify and rectify biases while providing feedback on biased outputs. Continuous monitoring is indispensable for detecting and rectifying biases that may emerge over time. Community engagement in data collection is crucial to ensure datasets capture marginalized communities' specific needs and challenges, contributing to the ethical and justice development and application of ML models in the energy sector. For example, community liaisons or representatives could be involved in designing a culturally appropriate survey with researchers and facilitating data collection efforts within their local communities. This approach helps ensure the data gathered reflects residents' lived experiences, energy usage patterns, and priorities rather than relying solely on external researchers' assumptions or limited perspectives. Community representatives could also aid researchers in accessing hard-to-reach populations and building trust to facilitate more comprehensive data gathering within those communities. This level of engagement helps center the voices and realities of marginalized groups in the data used to train ML models rather than having outside researchers make potentially biased assumptions. It promotes developing ethical, equitable models grounded in the actual needs of impacted communities. Researchers, practitioners, and stakeholders must establish guidelines to minimize biases and societal implications without a formal regulatory framework. This process involves the establishment of ethical guidelines and governance policies covering data privacy protection, ML system provenance documentation, acceptable bias thresholds, and a roadmap for bias mitigation [35,36].

Finally, given ML systems' uncertainties and data dependencies, transdisciplinary educational outreach aligned with the ethics and accountability tenets of the FATE framework is vital in mitigating biases that could adversely affect certain demographic groups. Institutions and stakeholders should encourage diversity in expertise within AI development and maintenance teams, integrating different disciplinary perspectives and including practitioners and decision-makers to uncover and address biases. This approach promotes accountability to affected groups and facilitates uncovering blind spots through diverse viewpoints, a core FATE principle.

User education and awareness efforts are essential, as users often have limited awareness of operational and application-related pitfalls. Developing best practices and educational programs grounded in FATE can help users interpret model predictions through an ethical lens, detect biases effectively, and uphold principles like privacy protection and informed consent. By fostering transdisciplinary collaboration and widespread education on ethical AI development in line with FATE, the energy sector can proactively institute processes prioritizing fairness, accountability, and justice in integrating machine learning technologies. The pursuit of bias reduction is a continuous effort; achieving complete bias elimination in any model remains an unattainable goal. The aim, however, is to consistently diminish bias, improve fairness within ML systems, and advocate for energy justice. By implementing these strategies, researchers can advance the development of ML models that actively support justice, equity, and sustainability within the energy sector. A sustained dedication to research and innovation in this field is crucial for effectively mitigating bias in energy-related ML applications.

5. Conclusions

As ML becomes essential for analyzing energy consumption data and energy efficiency, addressing data bias is vital to ensure energy justice and equitable access to energy services. ML can unintentionally perpetuate implicit biases that originate at various stages of design, development, and deployment, with significant ethical and societal implications. This perspective explores the complex relationship between data bias and energy justice, identifies different types of biases, and proposes strategies to mitigate these biases, thereby promoting fairness in ML applications in the energy sector. Effective measures include diversifying data sources, implementing bias mitigation techniques, enhancing transparency and accountability, and fostering interdisciplinary collaboration. By adopting these strategies, researchers and practitioners can create fairer and more ethical ML models, harnessing ML's potential to support a more equitable and sustainable energy future.

CRediT authorship contribution statement

Chien-fei Chen: Writing – original draft. **Bing Yao:** Writing – original draft. **Bandana Kar:** Writing – original draft. **Yuqing Hu:** Writing – original draft. **Rebecca Napolitano:** Writing – original draft.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Data availability

No data was used for the research described in the article.

Ackowledgements

Chien-fei Chen thanks Wellcome Trust Fundation, UK to support her work and the U.S. National Science Foundation under Grant SBE-2334298. Bandana Kar is funded by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Building Technologies Office, Award Number DE-EE00009748. The view expressed herein do not necessarily represent the view of the U.S. Department of Energy or the United States Government.

Rebecca Napolitano is funded by the U.S. National Science Foundation under Grant CMMI-2222849. Any opinions, findings, and conclusions or recommendations expressed in this material are thoes of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] L. Dencik, A. Hintz, J. Redden, E. Treré, Exploring data justice: conceptions, applications and directions. *Inf*, Commun. Soc. 22 (2019).
[2] K. Jenkins, D. McCauley, R. Heffron, H. Stephan, R. Rehner, Energy justice: a conceptual review, Energy Res. Soc. Sci. 11 (2016) 174–182.
[3] C. Chen, fei, Greig, J., Nelson, H. & Li, F., When disadvantage collides: the concentrated effects of energy insecurity and internet burdens in the United States. *Energy res*, Sociol. Sci. 91 (2022).
[4] B.K. Sovacool, M. Burke, L. Baker, C.K. Kotikalapudi, H. Wlokas, New Frontiers and conceptual frameworks for energy justice, Energy Policy 105 (2017) 677–691.
[5] T. Memmott, S. Carley, M. Graff, D.M. Konisky, Sociodemographic disparities in energy insecurity among low-income households before and during the COVID-19 pandemic, Nat. Energy 6 (2021) 186–193.
[6] United states department of energy. Building performance database; November 2023. <https://buildings.lbl.gov/cbs/bpd/.
[7] P.A. Mathew, et al., Big-data for building energy performance: lessons from assembling a very large National Database of building energy use, Appl. Energy 140 (2015) 85–93.
[8] H.A.U. Khan, S. Price, C. Avraam, Y. Dvorkin, Inequitable access to EV charging infrastructure, Electr. J. 35 (3) (2022) 107096, https://doi.org/10.1016/j.tej.2022.107096.
[9] U.S. Energy Information Administration Department of Energy. Residential Energy Consumption Survey (RECS) (2020).
[10] A. Bhattacherjee, Social science research: principles, methods, and practices, Book 3 (2012).
[11] C.M. Wade, J.S. Baker, G. Latta, S.B. Ohrel, Evaluating potential sources of aggregation Bias with a structural optimization model of the U.S. Forest sector, J. For. Econ. 34 (2019).
[12] Agency for Toxic Substances and Disease Registry. Social Vulnerability Index (SVI). https://www.atsdr.cdc.gov/placeandhealth/svi/index.html (2023).

[13] A.S. Fotheringham, D.W.S. Wong, The modifiable areal unit problem in multivariate statistical analysis, Environment and Planning A: Economy and Space 23 (7) (1991) 1025–1044.
[14] B. Kar, M.E. Hodgson, Observational scale and modeled potential residential loss from a storm surge, GISRS 49 (2012) 202–227.
[15] B. Khavari, A. Sahlberg, W. Usher, A. Korkovelos, F.F. Nerini, The effects of population aggregation in geospatial electrification planning, Energ. Strat. Rev. 38 (2021).
[16] T. Cheng, M. Adepeju, Modifiable temporal unit problem (MTUP) and its effect on space-time cluster detection, PLoS One 9 (6) (2014) e100465.
[17] D. Mitra, N. Steinmetz, Y. Chu, K.S. Cetin, Typical occupancy profiles and behaviors in residential buildings in the United States, Energ. Buildings 210 (2020) 109713.
[18] Y.-S. Chiou, K.M. Carley, C.I. Davidson, M.P. Johnson, A high spatial resolution residential energy model based on American time use survey data and the bootstrap sampling method, Energ. Buildings 43 (2011) 3528–3538.
[19] Z. Yi, B. Chen, X.C. Liu, R. Wei, J. Chen, Z. Chen, An agent-based modeling approach for public charging demand estimation and Charging Station location optimization at urban scale, *computers*, Environment and Urban Systems 101 (2023).
[20] L. Diao, Y. Sun, Z. Chen, J. Chen, Modeling energy consumption in residential buildings: a bottom-up analysis based on occupant behavior pattern clustering and stochastic simulation, Energ. Buildings 147 (2017) 47–66.
[21] G. Ruan, et al., A cross-domain approach to analyzing the short-run impact of COVID-19 on the U.S. Electricity Sector, SSRN Electron. J. 4 (11) (2020) 2322–2337.
[22] Doma, A. *et al.* Towards Post-pandemic Occupancy Patterns: Investigating Building Occupancy in New York City. in *ASHRAE Transactions* vol. 128 323–330 (ASHRAE, 2022).
[24] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on Bias and fairness in machine learning, ACM Comput. Surv. 54 (2021).
[25] W. Chen, Y. Qiu, Y. Feng, Y. Li, A. Kusiak, Diagnosis of wind turbine faults with transfer learning algorithms, Renew. Energy 163 (2021) 2053–2067.
[26] G. Menardi, N. Torelli, Training and assessing classification rules with imbalanced data, Data Min. Knowl. Disc. 28 (2014) 92–122.
[27] T. Darrell, et al., Machine learning with interdependent and non-identically distributed data, Dagstuhl Reports 5 (2015).
[28] S. Kapoor, A. Narayanan, Leakage and the reproducibility crisis in machine-learning-based science, Patterns 4 (9) (2023).
[29] R.A. Poldrack, G. Huckins, G. Varoquaux, Establishment of best practices for evidence for prediction: a review, JAMA Psychiatry 77 (5) (2020) 534–540.
[30] J.M. Hofman, D.J. Watts, S. Athey, F. Garip, T.L. Griffiths, J. Kleinberg, H. Margetts, S. Mullainathan, M.J. Salganik, S. Vazire, A. Vespignani, T. Yarkoni, Integrating explanation and prediction in computational social science, Nature 595 (7866) (2021) 181–188.
[31] U.S. Energy Information Administration. Residential Energy Consumption Survey (RECS). (2022).
[32] C. Milchram, R. Hillerbrand, G. Kaa, N. Doorn, R. Künneke, Energy justice and smart grid systems: evidence from the Netherlands and the United Kingdom, Appl. Energy 229 (2018) 1244–1259.
[34] N.T. Lee, P. Resnick, & Barton, G, Best Practices and Policies to Reduce Consumer Harms, Algorithmic Bias Detection and Mitigation, 2019.
[35] M. Cannarsa, Ethics guidelines for trustworthy AI, The Cambridge Handbook of Lawyering in the Digital Age (2021), https://doi.org/10.1017/9781108936040.022.
[36] P. Nevels, Connected communities: a vision for the future of electric utilities, IEEE Rev. 48 (1) (2020).
[38] S. Bird, M. Dudick, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, K. Walker, Fairlearn: a toolkit for assessing and improving fairness in AI, Microsoft Tech. Rep. MSR-TR-2020-32 (2020).
[39] Google LLC. ML Fairness Gym (v 0.1.0). https://github.com/google/ml-fairness-gym (2024).
[40] IBM, AI Fairness 360 Toolkit, Open Source IBM, 2018.