January 22, 2025, Lecture Note: Data-Driven Systems

Sina Tayebati, Yashith Reddy Enamala

CS 516: Responsible Data Science and Algorithmic Fairness; Spring 2025
Abolfazl Asudeh; www.cs.uic.edu/∼ *asudeh/teaching/archive/cs*516*spring*25/

1 Introduction

This lecture centered on the pipeline of data-driven systems and the complexity of integrating fairness considerations within them. We will explore how data are collected, curated, and used in algorithmic decision-making, as well as how hidden biases and historical inequalities can compromise system outcomes. We will also discuss the interplay between purely mathematical or utilitarian perspectives on fairness and broader ethical or societal norms, and we will look at how algorithms trained on historical data can inadvertently propagate or amplify unwanted biases.

We begin with a series of classical moral dilemmas to illustrate how people's notions of fairness and justice can shift drastically with small changes in context, even if the numerical outcomes seem similar. These scenarios serve as a conceptual foundation, demonstrating that ethical judgments about fairness cannot be captured by single values or formulas. By the end of the lecture, you will gain an appreciation of why fairness must be viewed as a multifaceted construct that requires careful governance and auditing throughout the data lifecycle, rather than a single, static metric.

2 Ethical Foundations of Fairness

Discussions of fairness in data science must be rooted in moral philosophy and ethical theory, because purely mathematical definitions of fairness can hide complex normative questions about which outcomes are "just" or "equitable." When data scientists understand these foundational ideas, they are better equipped to examine algorithmic designs not solely by performance metrics, but also in light of larger social values and moral responsibilities.

2.1 Classical Moral Dilemmas

Moral dilemmas provide an effective way to illustrate how ethical and societal values can override simple cost-benefit calculations. Even when two situations appear numerically identical—such as weighing the sacrifice of one individual against the lives of many—small changes in the context, agent roles, or social norms can lead to drastically different judgments about what is morally acceptable. In this subsection, we present three classical thought experiments that highlight how our intuitions about "fairness" and "justice" can vary once we consider issues of directness of harm, individual rights, and social responsibilities.

Trolley Problem. In the Trolley Problem, a runaway trolley is heading down a track toward a group of people. Pulling a lever can redirect the trolley onto another track where it will kill only one person instead of several. On the surface, this scenario represents a straightforward numerical dilemma: is it justifiable to sacrifice one individual to save a larger group? However, deeper ethical questions arise concerning agency and responsibility. Does the act of pulling the lever establish direct moral culpability for the resulting harm, or is it merely a passive redirection of an already tragic event? Opinions differ on whether the lever-puller becomes morally liable for the death of the single person on the side track. Additionally, some argue that intentionally causing harm to one individual—however small in number—contradicts the principle that one should never actively harm someone else, even if it prevents a larger harm. Thus, while a simplistic utilitarian view might prioritize saving the greater number, other ethical frameworks emphasize the moral or legal implications of any act that intentionally causes a person's death, challenging the notion that sheer numbers should dictate the decision.

Island Riot Scenario. In the Island Riot Scenario, a remote village faces the threat of a riot that could harm or kill many inhabitants unless one innocent person is handed over or sacrificed to appease the rioters. Although it similarly presents a stark choice of sacrificing one to save many, the social and contextual factors are distinctly different from the Trolley Problem. Here, the directness of the harm—and the fact that a group of people is effectively forcing the sacrifice—raises complex questions about complicity and coercion. Some argue that succumbing to the demands of the rioters is morally objectionable because it rewards or legitimizes violent threats. Others might see it as a tragic necessity if it indeed prevents widespread harm. This dilemma illuminates how *intent* and *coercion* matter a great deal in moral reasoning, as we consider whether it is ever acceptable to place the welfare of a majority above the rights of an individual when the threat is actively imposed by other human agents rather than a natural or mechanical force.

Surgeon's Dilemma. The Surgeon's Dilemma describes a medical situation in which a surgeon could harvest organs from a healthy patient without their consent to save multiple other patients who need transplants. At first glance, this scenario echoes the numbers-oriented logic of the Trolley Problem: lose one person, save many. Yet it adds a decisive layer of ethical tension: a doctor's duty to do no harm and the strong social norm of respecting bodily autonomy. While purely utilitarian frameworks might treat the healthy patient's organs as resources to be leveraged for the greater good, other moral systems highlight the sanctity of individual rights, informed consent, and professional obligations. Indeed, forcibly taking organs without consent is widely considered a grave violation of medical ethics and personal autonomy. This highlights the role of specialized codes of conduct—for instance, in the medical profession—and the limits of purely numerical reasoning about harm.

These examples highlighted that purely utilitarian logic—maximizing total lives saved—does not always match public intuitions about fairness. Specific roles, norms, and taboos heavily influence moral choices. Moreover, we discussed the difference between *act utilitarianism* (evaluating each act's immediate consequences) and *rule utilitarianism* (adhering to broader rules or norms that maintain social trust). Although the Trolley Problem might look like a straightforward utility maximization scenario, organ harvesting is far more troubling even if it saves multiple lives. By extension, developers of data-driven systems must acknowledge that focusing purely on some metric of utility may clash with deeper ethical values—such as prohibitions against certain forms of discrimination.

3 Fairness in Algorithmic Decision-Making

After demonstrating that moral and ethical considerations guide our notions of fairness, we need to explain how these principles translate into data-driven algorithms. In practice, machine learning developers often articulate fairness constraints or regularizations to ensure that outcomes do not disproportionately harm certain groups. Nevertheless, such fairness definitions frequently conflict with each other, as illustrated by real-world examples.

3.1 Case Study: Recidivism Scores and COMPAS

A prominent illustration of fairness controversies in algorithmic decision-making can be found in the COMPAS tool, short for *Correctional Offender Management Profiling for Alternative Sanctions*. Developed by a company formerly known as Northpointe Inc. (now Equivant), COMPAS is used in various stages of the criminal justice process—such as setting bail or sentencing recommendations—to generate a numeric score estimating the likelihood that an individual will reoffend.

Despite its widespread adoption, COMPAS has been the focus of intense scrutiny after an investigative report by ProPublica accused it of exhibiting racially biased outcomes. Specifically, Black defendants who did not go on to commit new crimes were disproportionately classified as "high risk," indicating a false positive error. Conversely, White defendants who reoffended were more often labeled "low risk," corresponding to a false negative error. ProPublica thus concluded that the tool generated systematically unfair risk assessments.

In response, Northpointe defended the system by noting that COMPAS achieves similar *accuracy* and *predictive parity* across racial groups. Predictive parity means that, among those labeled "high risk," the proportion who actually reoffend is roughly the same for different demographic groups. Northpointe argued that this form of parity is a crucial indicator of fairness. Meanwhile,

United States courts have also weighed in on the matter, with the Wisconsin Supreme Court finding that the false positive/negative rates (or error rates) for different groups showed sufficient balance to justify the continued use of COMPAS. The Court thereby ruled in favor of allowing this risk assessment to influence judicial decisions.

The COMPAS case lays bare the complexities of defining "fairness" in predictive models. Although the tool can be shown to satisfy certain fairness metrics (like predictive parity or balanced error rates), it fails others (such as demographic parity, which would require equal proportions of "high risk" across groups). These discrepancies underscore the so-called "impossibility theorems," which note that, depending on the data and baseline rates of recidivism, it is not possible to fulfill all fairness criteria simultaneously. COMPAS thus exemplifies how real-world deployed systems confront the tension of multiple, and sometimes conflicting, fairness standards. The controversy also raises deeper ethical questions about the acceptability of algorithmic risk scores in high-stakes settings like the criminal justice system, where inaccurate labels can significantly impact an individual's future.

4 Data-Driven System Pipeline

Data is arguably the most important component in any data-driven system, as it fundamentally shapes the patterns and insights that machine learning models discern. In the context of responsible data science, understanding how data is gathered, curated, and used is crucial, because biases that originate in the data can carry through an entire pipeline and manifest in harmful ways during deployment.

One key consideration involves the inherent tension between various fairness measures, as highlighted by well-known impossibility theorems. These theorems show that not all fairness criteria (such as demographic parity, equal error rates, or predictive parity) can be satisfied simultaneously if there are baseline differences among demographic groups or if certain attributes (like race) correlate with the target variable. As seen in the COMPAS example, accuracy or error-rate parity alone may not suffice to achieve demographic parity; in other words, the dataset or underlying environment may not allow for simultaneously fulfilling every fairness goal. If the data itself is "biased"—for instance, if one group has historically been surveilled or policed more heavily—then a model built on that data will often reflect or even amplify such disparities.

Moreover, data, especially those capturing social processes, almost always contains historical biases and stereotypes. For instance, face detection algorithms that are trained predominantly on images of white engineers can fail to recognize darker-skinned faces. Selection biases also arise when the methods used to collect data do not produce samples that fully represent the broader population. These issues can cascade into a feedback loop, where biased data leads to biased decisions, which in turn produce further skewed data in the future.

Despite these challenges, data-driven approaches can also *reduce* human subjectivity if data is carefully collected and curated, thus potentially minimizing individual prejudices. The key is ensuring that data itself is managed responsibly. Researchers in database systems and machine learning have proposed various methods for addressing data bias, such as pre-processing the dataset to remove discriminatory correlations, employing causal reasoning to identify admissible variables, or ensuring adequate coverage for minority groups so that models are not trained exclusively on majority populations. However, this research remains at an early stage and requires a concerted effort from both the data management community and algorithmic fairness researchers.

In the remainder of this pipeline overview, we break down each stage—Data Collection, Curation, Model Training, Deployment, and Fairness Auditing—to illustrate where bias can enter and how responsible data practices can help mitigate it. These stages do not stand alone but interact with each other in complex ways. From correcting bias in the raw data, to designing fair representations, to integrating fairness into query processing and system architectures, each step calls for careful consideration of how seemingly technical choices can have significant social and ethical ramifications.

Data Collection. Real-world data may reflect historical inequalities, sampling biases, or incomplete records. For example, a mailed survey might ask "Do you enjoy taking surveys?"—a question almost guaranteed to be answered disproportionately by survey enthusiasts, thus skewing results.

Curation. Once collected, data must be cleaned, annotated, and organized. Labeling bias and assumptions can become embedded if not handled with care. If researchers rely on stereotypes (e.g., pink clothes $\rightarrow$ female), they risk amplifying cultural biases.

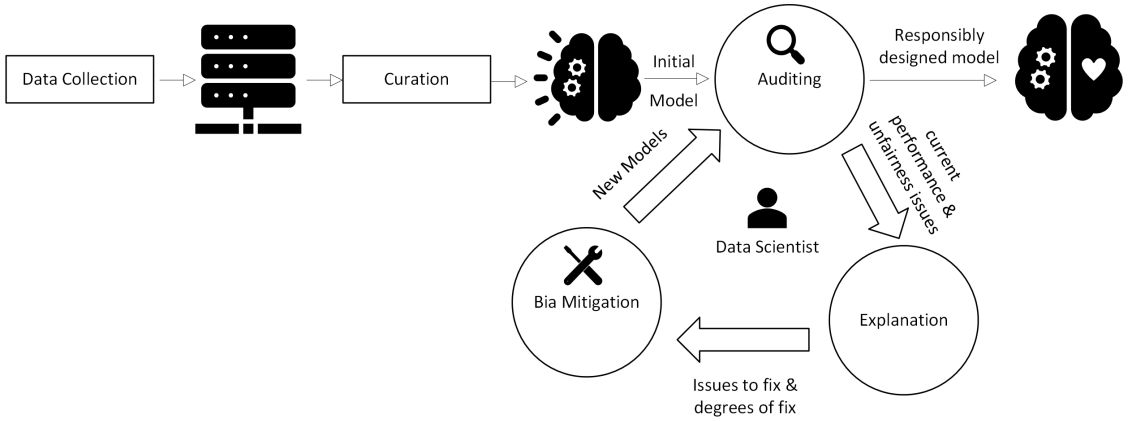

Figure 1: Processes and steps in designing a fair system [2].

Model Training. The learning algorithm minimizes a loss function, often focusing purely on predictive accuracy. However, if sensitive attributes (like race or gender) correlate with the target, the model may unwittingly use them—or correlated features like ZIP code—as proxies, leading to discrimination unless additional fairness constraints are introduced.
Deployment. Once a model is used in real-life settings, decisions made by the algorithm can have significant personal and societal effects (e.g., granting loans, setting bail, diagnosing patients). Biases that were dormant in a training set can manifest here.
Fairness Audit. Ongoing monitoring is essential. Even if a model is fair at launch, data distributions and user behaviors can shift over time, creating new patterns of unfairness. Fairness audits help detect these issues and allow for retraining or recalibration of the model.

5 Dataset Formalization and Sensitive Attributes

In machine learning and big data analytics, dataset formalization is essential for ensuring accuracy, fairness, and reliability. It involves structuring data in a way that enables meaningful analysis while addressing biases that may arise from sensitive attributes such as race, gender, age, and socioeconomic status. If improperly handled, these attributes can lead to discrimination in algorithmic decision-making. A dataset is mathematically represented as (x,s,y), where X represents non-sensitive features, s denotes sensitive attributes, and y is the target outcome.

Sensitive attributes can either be explicitly protected or indirectly influential through observational attributes. The following table illustrates this distinction

Table 1: Comparing Protected vs. Observational Attributes

	Protected	**Observational**
Examples	Race, Gender	Education, ZIP Code
Inference Usage	Restricted	Generally Permitted
Availability	Often Missing	Typically Available

Simply removing sensitive attributes does not eliminate bias, as observational attributes can act as proxies. A better approach is fairness-aware processing through three methods such as pre-processing, in-processing, and post-processing.

Bias amplification occurs when historical discrimination in data is learned and perpetuated by models. Additionally, intersectionality—the overlap of multiple protected attributes (e.g., race and gender)—can compound discrimination, requiring specialized mitigation strategies.

Causal inference techniques help distinguish between legitimate correlations and unjust biases, allowing models to adjust for unfair disparities. Proper dataset formalization is not only a technical requirement but an ethical necessity to ensure equitable AI systems.

6 Machine Learning Foundations and Fairness Constraints

Machine learning models aim to optimize predictive accuracy by minimizing a loss function

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{n} \mathcal{L}(h_\theta(x_i), y_i),$$

where h_θ represents the model's prediction and L quantifies the error. However, ensuring fairness requires introducing constraints that prevent bias against certain demographic groups.

Statistical parity ensures equal positive outcome probabilities across groups

$$\mathbb{P}[\hat{y} \mid s = 1] = \mathbb{P}[\hat{y} \mid s = 0]$$

Equalized odds ensures equal true and false positive rates across groups

$$\mathbb{P}[\hat{y} = 1 \mid y = 1, s = 1] = \mathbb{P}[\hat{y} = 1 \mid y = 1, S = 0]$$

$$\mathbb{P}[\hat{y} = 1 \mid y = 0, s = 1] = \mathbb{P}[\hat{y} = 1 \mid y = 0, s = 0]$$

Predictive parity ensures consistent accuracy across groups

$$\mathbb{P}[y = 1 \mid \hat{y = 1}, s = 1] = \mathbb{P}[y = 1 \mid \hat{y = 1}, s = 0]$$

Implementing fairness often leads to a fairness-accuracy trade-off, where strict fairness constraints may reduce model performance. Fairness aware learning can be categorized into pre-processing (modifying data), in-processing (integrating fairness into training), and post-processing (adjusting predictions) methods. Pre-processing methods adjust the data before training, in-processing methods integrate fairness constraints directly into model training, and post-processing methods adjust model predictions after training to ensure fairness. As machine learning models are deployed in high-stakes environments, continuous research and the integration of fairness constraints remain critical to avoiding systemic discrimination and ensuring equitable outcomes for all users.

Balancing fairness with predictive accuracy presents a key challenge, as enforcing fairness constraints can reduce model performance. To address this, techniques such as regularization, adversarial debiasing, and constrained optimization have been developed to maintain a reasonable level of accuracy while improving fairness. These methods help navigate the trade-off between equity and utility in machine learning models.

7 Practical Challenges in Fair Machine Learning

Despite advancements in fairness-aware machine learning, several practical challenges remain, stemming from societal biases, data limitations, regulatory constraints, and the trade-offs between fairness and accuracy. One key challenge is bias propagation, where historical biases in datasets are carried over into model predictions, perpetuating systemic inequalities in areas like hiring and policing. Identifying and mitigating these biases without distorting legitimate patterns is a complex issue.

Another challenge is the lack of a universally accepted definition of fairness, leading to conflicts between fairness metrics in different contexts. For example, balancing demographic parity with equalized odds often proves impossible, forcing practitioners to make difficult trade-offs, especially in high-stakes domains like healthcare and finance. Additionally, the phenomenon of "unknown unknowns" occurs when models inadvertently learn bias from proxies such as ZIP codes or education history, making it harder to detect unfair outcomes.

Feedback loops, where biased decisions reinforce existing disparities, further complicate fairness efforts. For example, biased loan approvals can perpetuate financial inequality by denying credit to certain groups, while biased criminal justice tools can lead to over-policing. Overcoming these feedback loops requires proactive intervention and continuous fairness monitoring. Another obstacle is the fairness-accuracy trade-off, where enforcing fairness constraints may slightly reduce model performance, particularly in industries where accuracy is critical.

Finally, regulatory compliance adds complexity, as different regions have varying laws governing algorithmic decision-making. While regulations like GDPR and the Equal Credit Opportunity Act mandate transparency and anti-discrimination measures, they often lack clear technical guidance, leaving organizations to navigate a complex legal landscape. Addressing these challenges requires a combination of technical solutions, ethical considerations, and regulatory oversight, along with interdisciplinary collaboration to create AI systems that promote fairness and equity.

8 Ethical Implementation Framework

The ethical implementation framework for fair machine learning ensures that fairness is integral throughout the model development and deployment lifecycle. It includes pre-processing, in-processing, and post-processing techniques designed to address bias at different stages. Pre-processing methods, such as reweighting, data augmentation, and causal inference, are used to create more balanced datasets by addressing biases before training begins. This helps mitigate discrimination against underrepresented groups and ensures data is representative of diverse demographic populations.

In-processing techniques incorporate fairness constraints directly into the model's learning process. Approaches like adversarial debiasing and fairness regularization modify the model's learning algorithm to minimize bias dynamically as it learns, ensuring that fairness is maintained during model training. These techniques focus on preventing biased outcomes from emerging during the learning phase, rather than correcting them afterward.

Post-processing techniques, applied after model training, fine-tune the model's predictions to ensure equitable outcomes. These include thresholding methods to equalize performance metrics across demographic groups and calibration methods to align predicted probabilities. Such adjustments ensure fair and consistent predictions, especially in high-stakes decisions like hiring, lending, or criminal justice.

Continuous monitoring, human oversight, and regulatory compliance are crucial in maintaining fairness post-deployment. Fairness audits track model performance over time to detect and correct emerging biases, while transparency tools and human intervention ensure accountability. Additionally, adherence to legal frameworks, such as the GDPR or the Equal Credit Opportunity Act, ensures that models meet ethical and legal standards for fairness. This cyclical approach promotes long-term fairness and aligns AI-driven decision-making with societal values.

9 Open Problem and Research Directions

Fair machine learning still faces numerous open challenges that require further research and refinement. One major issue is the quantification of long-term societal impacts. Current fairness metrics often evaluate bias at a single point in time but fail to capture how algorithmic decisions influence systemic inequalities over extended periods. For example, biased hiring algorithms can reinforce economic disparities over generations. Researchers are working on dynamic fairness metrics that account for these cascading effects, but developing comprehensive solutions remains complex.

Another critical challenge is detecting and mitigating indirect discrimination, where models use proxy variables that correlate with sensitive attributes, leading to biased outcomes. Even when race or gender is removed from the dataset, features like ZIP code or educational background can still serve as proxies, making fairness harder to enforce. Adversarial debiasing and counterfactual fairness techniques are promising approaches, but they require further optimization to be computationally efficient and scalable for large datasets.

Fairness under concept drift is another unresolved problem, as models trained on past data may become biased over time due to evolving societal and economic conditions. Traditional fairness-aware algorithms assume static data distributions, making them ill-equipped to adapt to changes. Future research must focus on developing adaptive fairness mechanisms that continuously monitor and adjust models in response to shifting data patterns, ensuring long-term fairness in automated decision-making systems.

The trade-off between fairness and explainability also presents a challenge. Many fairness interventions, such as deep learning-based bias mitigation techniques, make AI models more complex and less interpretable. Regulators and stakeholders increasingly demand transparency in automated decision-making, but balancing fairness and interpretability remains difficult. Research into

interpretable fairness methods, such as explainable AI (XAI) and rule-based fairness adjustments, is crucial for ensuring accountability in AI-driven systems.

Finally, ensuring fairness in multi-stakeholder environments adds another layer of complexity. Different groups, including governments, private companies, and advocacy organizations, have varying definitions of fairness, making it difficult to establish universal standards. Moreover, privacy concerns limit access to demographic data needed for fairness audits. Future research should focus on privacy-preserving fairness techniques, participatory AI frameworks, and legal-compliant methods to ensure fairness across different domains while protecting sensitive user information.

10 Conclusion

The lecture concluded by underscoring that the design of responsible data-driven systems must be an interdisciplinary effort. While technical solutions—such as fairness constraints, debiasing methods, and interpretability toolkits—are vital, they cannot fully resolve underlying social inequities. Contextual knowledge, legal considerations, and ethical reasoning are equally necessary to ensure that the system aligns with societal values and does not inadvertently harm vulnerable communities.

Moving forward, the course will delve deeper into specialized fairness algorithms, data governance strategies, and frameworks for auditing. By examining both their strengths and limitations, students will develop a more comprehensive view of what it means to build, maintain, and govern data-driven systems responsibly.

References

[1] A. Asudeh. *Enabling Responsible Data Science in Practice.* Adapted from conference presentations and articles on data curation, algorithmic fairness, and practical fairness toolkits.

[2] H. Zhang, N. Shahbazi, X. Chu, and A. Asudeh. *FairRover: Explorative Model Building for Fair and Responsible Machine Learning.* In *Proceedings of the Fifth Workshop on Data Management for End-To-End Machine Learning*, pp. 1–10, 2021.

[3] C. O'Neil. *Excerpts on Weapons of Math Destruction.* Discusses feedback loops, historic biases, and the risks of large-scale algorithmic systems.