Research Article

Journal of Human Resource Sustainability and Organizational Studies

From Statistical Fairness to Epistemic Fairness: The DBSD Framework for AI Bias Mitigation

Yair Oppenheim*

Ph. D. of Tel Aviv University, The Lester and Sally Antin Faculty of Humanities, School of Philosophy, Linguistics and Science Studies, Israel

***Corresponding Author**
Yair Oppenheim, Ph. D. of Tel Aviv University, The Lester and Sally Antin Faculty of Humanities, School of Philosophy, Linguistics and Science Studies, Israel.

Submitted: 2026, May 25; **Accepted:** 2026, Jun 15; **Published:** 2026, Jun 26

Citation: Oppenheim, Y. (2026). From Statistical Fairness to Epistemic Fairness: The DBSD Framework for AI Bias Mitigation. *J Hum Res Sus Org Stud, 1*(1), 01-18.

Abstract

Artificial Intelligence (AI), Machine Learning (ML), Large Language Models (LLMs), and data-driven decision systems increasingly influence critical domains including hiring, lending, healthcare, insurance, criminal justice, and digital governance. However, contemporary AI systems frequently reproduce or amplify social inequalities through hidden semantic inference processes that operate beyond explicit discriminatory rules or observable prediction outputs. This article introduces the Deep Bias Systematic Deviation (DBSD) framework, a novel semantic-inferential model that reconceptualizes algorithmic bias as an asymmetric inference-flow phenomenon operating across semantic, behavioral, and networked information structures. Unlike conventional fairness approaches that treat bias primarily as unequal outputs, DBSD models bias as a function of knowledge pressure, semantic alignment, stereotype activation, inference current, and epistemic resistance.

The article reviews five major fairness metrics widely used in AI governance and regulatory auditing: Disparate Impact, Statistical Parity, Equal Opportunity, Equalized Odds, and Calibration Bias. For each metric, the article presents detailed numerical examples illustrating both the biased state and the mitigation process achieved through DBSD-based semantic correction mechanisms. The originality of the article lies in its transition from prediction-level fairness toward epistemic fairness. Instead of correcting only final decisions, DBSD regulates the semantic and inferential processes that generate those decisions. Consequently, the framework provides a scalable and theoretically grounded foundation for future research in AI fairness, semantic auditing, inference governance, and algorithmic accountability.

Keywords: Artificial Intelligence (AI), Algorithmic Bias, DBSD, AI Fairness, Epistemic Fairness, Semantic Alignment, Inference Flow, Embedding Bias, Explainable AI, Calibration Bias, Equalized Odds, Equal Opportunity, Statistical Parity, Disparate Impact, Bias Mitigation, AI Governance, Semantic Regulation, Inference Auditing, Large Language Models (LLMs)

1. Introduction

increasingly influences high-stakes domains including hiring, lending, healthcare, criminal justice, insurance, and digital governance. As these systems become deeply integrated into social and institutional decision-making processes, growing concern has emerged regarding their capacity to reproduce, amplify, or institutionalize social bias and discriminatory outcomes [1–4]. Traditional approaches to algorithmic fairness often conceptualize bias as unequal statistical outcomes, discriminatory prediction behavior, or disparate treatment across protected demographic groups [5–8]. However, contemporary AI systems frequently generate harmful asymmetries through latent semantic inference processes that operate beyond explicit decision rules, visible attributes, or formally encoded discriminatory criteria [1,9–11]. While traditional definitions successfully capture bias as prejudice, deviation, or discriminatory treatment, they do not fully explain

the inferential and epistemic structures [12] through which modern AI systems generate asymmetric knowledge about individuals and groups. In large-scale semantic systems, discrimination may emerge even in the absence of explicitly encoded discriminatory rules, because latent representations, embedding structures, proxy variables, and semantic associations enable the system to infer sensitive attributes indirectly [8,10,13,15-16]. This problem reflects a broader form of epistemic asymmetry [12] in which certain groups become disproportionately exposed to profiling, semantic categorization, and stereotype activation. In this sense, algorithmic bias increasingly operates not merely as statistical unfairness, but as a form of inferential and epistemic inequality.

The **Deep Bias Systematic Deviation (DBSD)** framework introduces a novel paradigm in which bias is reconceptualized not primarily as intentional discrimination, but as susceptibility to inference-driven deviation within contemporary information systems. Traditional bias models often focus on unequal outputs, historical prejudice or disparate treatment [21,22]. However, in the age of artificial intelligence and large language models, systematic bias can emerge even when no explicit discriminatory rule is encoded [23,24]. Inspired by information-flow theory, semantic inference analysis, and network-based models of knowledge propagation [18]. The DBSD framework conceptualizes **epistemic impedance** - the degree of resistance that prevents systems from transforming available signals into harmful asymmetric judgments about individuals or groups [19,20]. In formal terms, bias risk depends not only on which variables are explicitly used, but on how easily hidden attributes, preferences, health conditions, identities, or vulnerabilities can be inferred from semantic, behavioral, or relational traces and converted into unequal outcomes [25,13, 26]. By focusing on the dynamics of knowledge extraction rather than merely visible decisions [21, 22], the DBSD framework provides a new theoretical and practical foundation for bias engineering, fairness analysis, and regulatory governance in AI-driven environments. More broadly, future fairness analysis must move beyond static output auditing toward dynamic inference auditing, where the central question is no longer only what AI systems decide, but what forms of knowledge they are structurally capable of generating, propagating, and amplifying about individuals and social groups [6,21].

1.1. Contributions of This Work

The present article reviews five commonly used metrics for measuring the existence and magnitude of bias in AI and algorithmic systems. For each type of bias, the article presents: a numerical example illustrating the bias before correction using the DBSD framework, and a demonstration of how the bias can be mitigated under each metric through DBSD- based intervention mechanisms. In addition, each case study is accompanied by a detailed numerical example showing: the original biased state, the semantic and inferential correction process, and the resulting reduction in bias after applying the proposed framework.

1.2. The Research Contribution of the Present Article is Fourfold.

First, the article establishes a unified analytical framework connecting classical fairness metrics with the DBSD semantic-inferential perspective. While traditional fairness literature typically focuses on prediction outputs and statistical disparities, the proposed framework shifts the analysis toward the deeper inferential and semantic mechanisms that generate biased outcomes. Second, the article introduces a novel interpretation of bias as an asymmetric inference-flow phenomenon. Instead of treating bias solely as unequal prediction behavior, the framework models bias as a function of semantic alignment, stereotype activation, inference current, and epistemic resistance. This creates a new bridge between: AI fairness, semantic representation theory, and epistemic regulation. Third, the article demonstrates mathematically and numerically how semantic correction mechanisms can mitigate multiple fairness metrics simultaneously, including: Disparate Impact, Statistical Parity, Equal Opportunity, Equalized Odds, and Calibration Bias.

Unlike many traditional fairness approaches that rely primarily on post-processing or threshold balancing, the proposed DBSD approach intervenes earlier — at the semantic representation and inference generation stages. Fourth, the article proposes a regulatory-oriented framework capable of quantifying discriminatory inference asymmetry through concepts such as: inference current, bias impedance, and semantic resistance. This contribution may support future development of: explainable fairness auditing, AI governance mechanisms, semantic fairness regulation, and epistemic accountability standards for advanced AI systems. Overall, the article contributes a novel interdisciplinary framework that integrates: AI fairness, semantic embeddings, inference dynamics, critical theory, and epistemic analysis into a unified mathematical and conceptual model for understanding and mitigating algorithmic bias. The transition from statistical fairness toward epistemic fairness reflects a broader transformation in contemporary AI governance. As modern AI systems increasingly rely on latent semantic representations, distributed inference architectures, and large-scale behavioral profiling, regulating outputs alone becomes insufficient. The central challenge is no longer merely whether decisions are statistically balanced, but whether the inferential structures generating those decisions remain epistemically justifiable [12], transparent, and resistant to discriminatory semantic propagation.

2. Related Work: Regulation, Bias Detection, and Algorithmic Fairness

The rapid deployment of Artificial Intelligence (AI), Machine Learning (ML), Large Language Models (LLMs), and algorithmic decision-making systems has intensified global concerns regarding bias, discrimination, fairness, and transparency in digital environments [1–4]. Regulatory institutions, governmental agencies, and academic researchers increasingly recognize that algorithmic systems may reproduce, amplify, or institutionalize social inequalities embedded within training data, organizational processes, and networked information ecosystems [7,8]. Traditional

research on algorithmic bias primarily focused on statistical discrimination and fairness metrics within structured datasets [5]. Early approaches introduced formal fairness notions such as demographic parity, equal opportunity, equalized odds, predictive parity, and calibration-based fairness constraints [5,11,28]. These models attempted to mathematically quantify discriminatory outcomes across protected groups including gender, race, ethnicity, age, religion, and disability status. However, subsequent research demonstrated that many fairness metrics are mutually incompatible under realistic conditions, creating unavoidable trade-offs between competing fairness objectives [5,6].

In parallel, regulatory frameworks evolved from general anti-discrimination doctrines toward AI-specific governance models. Within the European Union, the proposed European Union AI Act establishes a risk-based framework for AI systems, emphasizing transparency, accountability, human oversight, bias mitigation, and conformity assessment procedures for high-risk systems [25,29]. The AI Act is complemented by existing legal instruments such as the European Commission General Data Protection Regulation (GDPR), the Digital Services Act (DSA), and the Charter of Fundamental Rights of the European Union, all of which indirectly influence algorithmic fairness and non-discrimination obligations [7,21,41]. In the United States, regulatory oversight is more fragmented and sector specific. Agencies such as the Federal Trade Commission (FTC), the Equal Employment Opportunity Commission (EEOC), the Consumer Financial Protection Bureau (CFPB), and the National Institute of Standards and Technology (NIST) have issued guidelines addressing algorithmic discrimination, automated decision systems, explainability, and fairness auditing [4, 22,36]. NIST's AI Risk Management Framework (AI RMF) particularly emphasizes governance, measurement, monitoring, and mitigation of harmful algorithmic behaviors [36].

Academic literature has increasingly explored the intersection between regulation and technical fairness measurement. Researchers proposed benchmark frameworks for evaluating whether algorithmic systems comply with regulatory fairness expectations [37,38]. Such studies commonly compare legal concepts of discrimination with statistical fairness metrics, explainability models, and automated auditing methodologies. Several works highlighted the growing gap between legal definitions of discrimination and the operational behavior of modern AI systems, particularly deep learning architectures and LLMs whose internal inference mechanisms remain partially opaque [39–41]. Recent research further demonstrated that bias cannot be reduced solely to observable outputs. Instead, bias emerges through complex inference pathways distributed across large-scale information networks [7,21]. Embedding representations, semantic similarity structures, recommendation systems, and data-fusion processes can generate latent discriminatory inferences even when explicit protected attributes are removed from datasets [42–44]. Consequently, contemporary studies increasingly investigate hidden bias propagation, representational harms, embedding-space discrimination, and systemic amplification mechanisms within interconnected digital ecosystems.

To address these challenges, researchers and regulators have begun integrating AI-based bias detection tools into auditing pipelines. Natural Language Processing (NLP) systems, transformer-based classifiers, semantic embedding analysis, explainability models such as SHAP and LIME, and fairness auditing platforms are now employed to identify biased language, discriminatory recommendations, and unequal decision outcomes. However, existing approaches still largely focus on local statistical properties or isolated model behaviors rather than on the broader network dynamics governing inference propagation and knowledge diffusion.

The present work differs fundamentally from prior approaches in several important respects.

First, existing regulatory and fairness frameworks generally conceptualize bias as a property of datasets, outputs, or model decisions. In contrast, this paper treats bias as a dynamic epistemic phenomenon emerging through distributed inference processes across interconnected information systems.

Second, while current fairness metrics evaluate observable disparities between protected groups, the proposed framework models the systemic resistance to discriminatory inference using a network-impedance perspective inspired by physical diffusion systems. This enables analysis not only of explicit discrimination but also of latent inference pathways and hidden semantic propagation.

Third, current regulatory mechanisms primarily rely on static compliance assessment, post-hoc auditing, or local fairness measurements. The proposed framework introduces a dynamic and scalable perspective capable of modeling bias propagation, semantic alignment, inference amplification, and regulatory intervention within continuously evolving AI ecosystems.

Finally, whereas prior work typically separates technical fairness analysis from regulatory theory, the proposed model attempts to bridge these domains by providing a unified mathematical framework connecting inference dynamics, semantic representations, network diffusion, and regulatory control mechanisms. In doing so, the framework contributes toward the development of measurable, operational, and theoretically grounded approaches for future AI governance and algorithmic accountability. Existing fairness frameworks significantly advanced the measurement and regulation of discriminatory outcomes in AI systems. However, most current approaches remain primarily output-oriented and insufficiently capture the semantic and inferential dynamics through which modern AI systems generate asymmetric knowledge about individuals and groups. This limitation motivates the transition from statistical fairness toward epistemic fairness [12] proposed in the present work.

3. Bias Fairness Metrics and Regulatory Audit Frameworks

The Meaning of the Symbol $\hat{Y}$

The Basic Meaning of $\hat{Y}$

Symbol meaning: Y – Ground truth / true outcome, $\hat{Y}$ – Model prediction

Simple Example

Suppose an AI model decides whether to approve a loan Table 1.

Person	True Outcome Y	Model Prediction Yˆ
Person 1	1 (will repay loan)	1
Person 2	1	0
Person 3	0	1
Person 4	0	0
Values Meaning: 1-Positive outcome, 0-Negative outcome		

Table 1

3.1. Disparate Impact (DI) [31,32]

Formula: DI = P(Ŷ = 1 | A= minority) / P(Ŷ = 1 | A = majority).

Regulatory threshold: DI < 0.8 may indicate discrimination [8, 42].

Datasets : hiring, loans, insurance.

3.1.1. Symbol Meaning: $Y = 1$ Candidate is truly qualified, $\hat{Y} = 1$ The algorithm recommends

Example: Majority approval = 80%, Minority approval = 50%, DI = 0.625.

3.2. Statistical Parity [7, 42]

Formula: P(Ŷ = 1 | A = a) = P(Ŷ = 1 | A = b).

Measures equal distribution of positive outcomes.

Statistical Parity asks:

Does the model give positive outcomes at the same rate to different groups?

Importantly: It does **not** check whether the prediction is correct. It only checks the distribution of positive decisions [7].

Example — Hiring System

Suppose an AI hiring system evaluates job applications. We have two groups: Group A = Men, Group B = Women

3.2.1. Model Decisions

Men: Total Applicants = 100, Approved by AI ($\hat{Y}$=1) = 80 **Women:** Total Applicants = 100, Approved by AI ($\hat{Y}$=1) = 40

Conclusion: Statistical Parity is violated.

3.2.2. Important Observation

Statistical Parity does **not** ask whether applicants were truly qualified.

It only checks: Are positive decisions distributed equally? This is one of the main criticisms of Statistical Parity [5,6,29].

Statistical Parity is useful when society wants equal access, demographic representation, inclusion, or anti-discrimination guarantees.

Common domains: hiring, university admissions, advertising, loan approvals.

3.2.3. Main Limitation

Suppose: Group A truly has 90 qualified candidates, Group B truly has 30 qualified candidates. Therefore, Forcing equal approval rates may: reduce model accuracy, create artificial balancing, or conflict with merit-based evaluation.

This is why Statistical Parity often conflicts with: Equal Opportunity, Equalized Odds, Calibration.

Datasets: hiring, admissions, advertising.

3.3. Equal Opportunity

Equal Opportunity focuses on equality in True Positive Rates [5, 43]

Formula: P(Ŷ = 1 | Y = 1, A = a) = P(Ŷ = 1 | Y = 1, A = b).

Measures equality in True Positive Rate. The person truly deserves a positive outcome, $\hat{Y} = 1$ the model gives a positive prediction, $A \rightarrow$ protected group attribute

3.3.1. Intuition

Equal Opportunity asks: Among people who truly deserve a positive outcome,

does the model treat all groups equally? This metric focuses on:" True Positive Rate (TPR)"

Example — Loan Approval System

Suppose an AI system decides whether to approve loans.

We examine only applicants who: truly would repay the loan,

that is: $Y = 1$, We compare: Men , Women

3.3.2. Ground Truth Data

Men: Truly qualified applicants (Y=1) =100, Approved by AI ($\hat{Y}$ = 1) = 90

Women: Truly qualified applicants (Y=1) = 100, Approved by AI ($\hat{Y}$ =1) =80

Step 1 — Compute True Positive Rates (TPR)
For men: $P(\hat{Y} = 1 | Y = 1, A = \text{"Men"}) = \frac{90}{100} = 0.90$
For women: $P(\hat{Y} = 1 | Y = 1, A = \text{"Women"}) = \frac{90}{100} = 0.80$
Since:0.95≠0.80 Equal Opportunity is violated.

3.3.3. Interpretation

The system is unfair because: qualified men receive loans much more often than qualified women. This means: one group experiences more False Negatives.

This metric is especially useful when: denying qualified individuals is highly harmful [5,21]. Examples: loan approvals, university admissions, medical treatment, hiring systems.

3.3.4. Differences from Statistical Parity

Statistical Parity ignores whether people are truly qualified.
Equal Opportunity considers only: $Y = 1$. Thus: Statistical Parity focuses on equal outcomes, Equal Opportunity focuses on equal treatment of qualified individuals.

3.3.5. Important Limitation

Equal Opportunity only equalizes: " PR". It does **not** constrain: False Positive Rates, Calibration, overall prediction rates. Therefore: a system may satisfy Equal Opportunity
while still discriminating in other ways.
Datasets : medicine, credit scoring.

3.4. Equalized Odds

Equalized Odds is one of the strongest fairness criteria [5,43,45]
Requires equal TPR and FPR across groups.
Equalized Odds requires: $P(\hat{Y} = 1 \mid Y = y, A = a) = P(\hat{Y} = 1 \mid Y = y, A = b)\ \forall\ y \in \{0,1\}$ This means: The model must have equal: True Positive Rates (TPR) , False Positive Rates (FPR) across groups.

3.4.1. Intuition Equalized Odds asks: Does the model make both kinds of errors equally across groups? Unlike Statistical Parity, Equalized Odds consider: Ground truth Y, Correct predictions, Errors

Example — Loan Approval System

Suppose an AI system decides whether to approve loans. Protected groups: Group A = Men, Group B = Women
Definitions: Table 2

Symbol	Meaning
Y = 1	Applicants truly deserve loan
Y = 0	Applicants truly should not receive loan
$\hat{Y} = 1$	Model approved loan
$\hat{Y} = 0$	Model denied loan

Table 2

Group A (Men)			Group B (Women)		
Reality Y	**Prediction $\hat{Y}$**	**Count**	**Reality Y**	**Prediction $\hat{Y}$**	**Count**
1	1	80	1	1	80
1	0	20	1	0	20
0	1	10	0	1	10
0	0	90	0	0	90

Table 3

Step 1 — Compute TPR for Men
True Positive Rate: $TPR_A = P(\hat{Y} = 1 | Y = 1, A = A)$ Total truly qualified men: 80 + 20 = 100, Correctly approved: 80, Therefore: $TPR_A = \frac{80}{100} = 0.80$

Step 2 — Compute FPR for Men False Positive Rate: $FPR_A = P(\hat{Y} = 1 \mid Y = 0, A = A)$ Total truly unqualified men:10 + 90 = 100, Incorrectly approved:10, Therefore: $FPR_A = \frac{10}{100} = 0.10$

Step 3 — Compute TPR for Women TPR_B = $\frac{80}{100} = 0.80$

Step 4 — Compute FPR for Women FPR_B= $\frac{10}{100} = 0.10$

Step 5 — Compare the Rates
We obtain: $TPR_A = TPR_B = 0.80$, and: $FPR_A = FPR_B = 0.10$

Conclusion
Equalized Odds is satisfied.
The model: approves qualified individuals equally, and makes false approvals equally.

3.4.2. Example of Violation

Suppose instead: Table 4

Men	Women				
Reality Y	**Prediction $\hat{Y}$**	**Count**	**Reality Y**	**Prediction $\hat{Y}$**	**Count**
1	1	90	1	1	70
1	0	10	1	0	30
0	1	5	0	1	25
0	0	95	0	0	75

Table 4

Compute TPR

Men: $TPR_A = \frac{90}{100} = 0.90$,Women: $TPR_B = \frac{70}{100} = 0.70$

Compute FPR

Men: $FPR_A = \frac{5}{100} = 0.05$,Women: $FPR_B = \frac{25}{100} = 0.25$

3.4.3. Interpretation

The model: correctly approves qualified men much more often and incorrectly approves unqualified women much more often.
Thus: Equalized Odds violated.

3.4.4. Why Equalized Odds Is Important

False Positives = approving risky borrowers, False Negatives = denying qualified borrowers Equalized Odds is considered one of the strongest fairness criteria because it controls: benefits, harms, error asymmetry [21, 43]. It is especially important in: criminal justice, medicine, fraud detection, facial recognition, and lending.

3.4.5. Main Limitation

Equalized Odds can: reduce accuracy, conflict with Calibration, become impossible under different base rates. This is one reason fairness metrics often contradict each other.

3.4.6. Famous Theoretical Result [44,45]

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan showed that:
Calibration and Equalized Odds cannot g 7enerally hold both when groups have different base rates. This became one of the foundational impossibility results in AI fairness.
Datasets : criminal justice, fraud detection.

3.5. Calibration

Calibration is crucial whenever probabilities drive decisions [6, 21, 36]
Calibration requires: $P(Y = 1 \mid \hat{Y} = p, A = a) = p$ where: $Y = 1 \rightarrow$ the event truly occurs, $\hat{Y} = p \rightarrow$ the model predicts probability p, $A \rightarrow$ protected group attribute

3.5.1. Intuition

Calibration asks: When the model predicts a probability of 80%, does the event actually happen about 80% of the time? A calibrated model produces probabilities that correspond to real-world frequencies.

Example — Loan Default Prediction Suppose a bank AI predicts: $\hat{Y} = 0.80$ meaning: "This applicant has an 80% probability of repaying the loan." We examine: Men, Women

Example of Calibration Failure

Suppose:
Men - Model predicts: $\hat{Y} = 0.80$ for 100 men. Observed: Truly repaid = 80, Defaulted = 20. Thus: $P(Y = 1 \mid \hat{Y} = 0.80, A = \text{"Men"}) = 0.80$
Women - Model predicts: $\hat{Y} = 0.80$ for 100 women.
Observed: Truly repaid = 50, Defaulted = 50. Thus: $P(Y = 1 \mid \hat{Y} = 0.80, A = \text{"Women"}) = 0.50$

3.5.2. Interpretation

The model says: "80% repayment probability"
But reality: only 50% repay. Thus: not calibrated for women.

3.5.3. Why Calibration Is Important

Calibration is crucial whenever: probabilities drive decisions, risk estimation matters, or uncertainty must be trusted.

Important domains: finance, insurance, healthcare, criminal justice, autonomous systems.

3.5.4. Difference Between Calibration and Accuracy

A model may: be accurate overall, but poorly calibrated.
Example: always predicting 90%, even when only 60% succeed.

3.5.5. Difference Between Calibration and Equalized Odds

Calibration focuses on: $P(Y = 1 \mid \hat{Y} = p)$
Equalized Odds focus on: TPR equality, FPR equality.
These are different fairness goals.

3.5.6. Famous Fairness Contradiction

One of the most important theoretical results in fairness research: Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan [6] proved that:
Calibration and Equalized Odds generally cannot hold both when groups have different base rates. This became one of the foundational impossibility theorems in AI fairness.

3.5.7. Simple Contradiction Example

Suppose: Men- True repayment rate 80%, Women True repayment rate 40% A calibrated model must reflect higher probabilities for men, lower probabilities for women. But then: TPR/FPR often becomes unequal.

Thus: Calibration satisfied, Equalized Odds violated.

3.5.8. Calibration Became Central in COMPAS Debate In the

COMPAS controversy [28, 43]. ProPublica emphasized unequal error rates, Northpointe emphasized calibration. This became the classic example showing different fairness metrics produce different conclusions.

4. Core Inferential Theory of DBSD

The Deep Bias Systematic Deviation (DBSD) framework conceptualizes algorithmic bias not merely as a statistical imbalance in outputs, but as a dynamic inferential process through which AI systems generate asymmetric knowledge about individuals and social groups. Traditional fairness approaches primarily evaluate prediction outcomes, classification parity, or error distributions. In contrast, DBSD shifts the analytical focus toward the semantic and inferential structures that produce those outcomes.

Inspired by information diffusion theory, network propagation models, semantic embedding analysis, and epistemic justice theory, DBSD interprets discriminatory bias as a form of asymmetric inferential exposure. Under this perspective, individuals and groups may become disproportionately vulnerable to semantic profiling, stereotype activation, and latent attribute inference even when explicit discriminatory rules are absent.

4.1. Core Conceptual Structure

The DBSD framework is based on four central conceptual components:

a. Semantic Alignment

Semantic alignment represents the degree to which an input embedding aligns with sensitive semantic prototypes associated with protected attributes, stereotypes, or latent identity categories [9,10,13,15,16,19,20].

b. Stereotype Activation

When semantic alignment becomes sufficiently strong, latent stereotype structures may become activated [10,13,24,36] within the model's inferential space, increasing the probability of discriminatory semantic inference.

c. Differential Inference Flow

DBSD models discriminatory behavior as unequal inferential flow across groups [18,21]. Certain groups may become disproportionately exposed to semantic inference, predictive amplification, or hidden profiling mechanisms.

d. Epistemic Impedance

Epistemic impedance represents the resistance [12, 18] that limits or constrains discriminatory inference generation. Higher impedance reduces semantic inference intensity and weakens downstream stereotype activation.

4.2. Mathematical Formulation

Let z denote an embedding vector representing an individual input and let c denote a sensitive semantic prototype. The semantic alignment between z and c is measured using cosine similarity [15, 16, 21]: cos(z,c) = (z · c) / (||z|| ||c||)

The probability that the system infers a sensitive attribute is modeled as: p = σ(β cos(z,c))

where:

- σ denotes the sigmoid activation function [15],
- β controls inferential sensitivity.

DBSD defines epistemic impedance as: Z = -log(p)

Let G_a and G_b be two social groups, for example men/women, majority/minority, young/old, or any protected/non-protected category.

For each group G, define: $I_B(G) = \frac{V_k(G)}{Z_B(G)}$

where:

- $I_B(G)$ = bias-relevant inference current for group G
- $V_k(G)$ = knowledge pressure applied to group G
- $Z_B(G)$ = bias impedance protecting group G from distorted, stereotypical, or discriminatory inference

Interpretation:

- High $V_k(G)$: stronger pressure to infer, classify, profile, rank, or predict members of group G
- Low $Z_B(G)$: weaker resistance to biased inference
- High $I_B(G)$: higher exposure to biased epistemic treatment

Under this formulation:

- stronger semantic alignment increases inference probability,
- higher inference probability lowers epistemic impedance,
- lower impedance increases discriminatory inference exposure.

4.2.1. Differential Bias Current

Bias between two groups is defined as the difference in inference current: $\Delta I_B(G_a, G_b) = |I_B(G_a) - I_B(G_b)|$ Substituting the impedance formulation: $\Delta I_B(G_a, G_b) = \left|\frac{V_k(G_a)}{Z_B(G_a)} - \frac{V_k(G_b)}{Z_B(G_b)}\right|$ This becomes the central DBSD-based bias metric.

4.2.2. Normalized Bias Flow Index

For regulatory and comparative purposes, define a normalized metric: $BFI(G_a, G_b) = \frac{|I_B(G_a) - I_B(G_b)|}{I_B(G_a) + I_B(G_b) + \epsilon}$ where $\epsilon > 0$ prevents division by zero. Thus: $0 \leq DBFI \leq 1$

DBFI value	Meaning
near 0	low differential bias flow
moderate	measurable asymmetric inference
near 1	severe bias imbalance

Interpretation: Table 5

4.3. Inferential Dynamics

Unlike traditional fairness models that evaluate isolated outputs, DBSD interprets AI systems as distributed inferential environments [18,23,24] in which semantic information propagates dynamically across interconnected representations.

Accordingly, discrimination may emerge not only from explicit classification rules, but also from cumulative semantic propagation [18,30] processes including:

- latent stereotype reinforcement,
- proxy attribute inference,
- semantic clustering,
- contextual amplification,
- cascading representational effects.

This inferential perspective aligns with network diffusion theory and information propagation models, where weak semantic signals may gradually amplify across interconnected structures.

4.4. Epistemic Fairness

The DBSD framework generalizes traditional statistical fairness into a broader concept of epistemic fairness [12]. Under this perspective, fairness is no longer evaluated solely by examining whether outputs are statistically balanced, but also by evaluating whether inferential processes themselves remain epistemically justifiable [12].

Consequently, the central analytical question shifts from:
"What decision did the model produce?" toward: "What forms of knowledge is the model structurally capable of generating about different individuals and groups?"
This transition from output-oriented fairness toward inferential and epistemic fairness constitutes the central theoretical contribution of the DBSD framework.

4.5. Theoretical Propositions

Proposition 1: Statistical fairness violations may emerge from asymmetric inferential exposure rather than solely from explicit discriminatory rules [8,21].
Proposition 2: Reducing semantic alignment with sensitive prototypes decreases discriminatory inference probability and weakens stereotype activation [9,13,32].
Proposition 3: Epistemic fairness generalizes output-based fairness metrics by regulating inferential structures in addition to observable predictions [12].
Proposition 4: Discriminatory semantic propagation may exhibit diffusion-like behavior across interconnected inferential systems [18,30].

5. DBSD Framework Could Mitigate Bias Fairness Metrics

While the five-fairness metrics reviewed in this article differ operationally, they share a common inferential substrate. DBSD assumes that disparate outcomes, unequal opportunity, calibration errors, and asymmetric error rates originate from underlying differences in semantic inference generation. Consequently, the framework analyzes all fairness metrics through a unified inferential lens grounded in semantic alignment, stereotype activation, inference current, and epistemic impedance [12,18,21]. Traditional fairness approaches attempt to mitigate every bias type by rebalancing datasets, removing protected attributes, imposing fairness constraints, or post-processing predictions. However, DBSD argues that these methods often address only: the final output layer. The deeper problem lies earlier: within the structure of inference generation itself.

5.1. DBSD Interpretation of Disparate Impact

Under DBSD, Disparate Impact is not merely: unequal decisions, but rather: unequal epistemic exposure, unequal inference pressure, and unequal semantic accessibility.

DBSD argues that the bias often emerges because vulnerable groups experience: higher monitoring, denser profiling, stronger semantic alignment, and lower inference resistance. For each group G, define: $I_B(G) = \frac{V_k(G)}{Z_B(G)}$. Thus: $I_P(G_{minority}) > I_P(G_{majority})$.

This produces more aggressive classification, more risk activation, more negative predictions, and eventually lower positive outcome rates. The following subsections apply the same DBSD inferential framework to different fairness metrics. Therefore, certain core concepts—including semantic alignment, inference current, and epistemic impedance—remain structurally similar across the examples.

5.1.1. DBSD Mitigation Strategy

Instead of merely forcing equal outputs, DBSD attempts to regulate: the inference flow itself. DBSD mitigates the bias by increasing: $Z_P(G_{minority})$ Possible mitigation mechanisms: semantic abstraction, embedding perturbation, inference throttling, controlled uncertainty, proxy suppression, feature decorrelation, stereotype attenuation.

5.1.2. Semantic Alignment Reduction

Suppose: z_i is a user embedding, and: c_G is a stereotype prototype vector. DBSD transforms: $z_i' = z_i - \gamma \frac{z_i \cdot c_G}{\|c_G\|^2} c_G$. This reduces excessive stereotype alignment.

5.1.3. Proposed DBSD-Based Regulatory Principal

A regulator may require: $\Delta IP\ (G_a, G_b) \leq \delta$ meaning: the difference in inference current between groups must remain below a defined threshold. This creates measurable epistemic fairness, rather than only statistical fairness [12].

5.1.4. Mitigating Disparate Impact Bias Using the DBSD Framework

Disparate Impact (DI) measures whether different social groups receive positive outcomes at substantially different rates.
Formally: $DI = \frac{P(\hat{Y}=1|A=minority)}{P(\hat{Y}=1|A=majority)}$, Under the traditional "80% rule": $DI < 0.8$
may indicate potential discrimination.

Typical examples: hiring systems, loan approvals, insurance pricing, admissions systems.

Traditional fairness approaches attempt to mitigate DI by rebalancing datasets, removing protected attributes, imposing fairness constraints, or post-processing predictions.

However, DBSD argues that these methods often address only: the final output layer.

The deeper problem lies earlier: within the structure of inference generation itself.

Example — Loan Approval Suppose: Majority - Positive Approval Rate = 80% , Minority - Positive Approval Rate = 40%, Then $DI = \frac{0.40}{0.80} = 0.50$. This violates the 80% rule.

Traditional fairness says: "the outputs are unequal." DBSD asks: Why did the inference process become asymmetrical?

5.1.5. DBSD Root-Cause Explanation

Suppose minority applicants are subjected to more behavioral profiling, more proxy-variable inference, more historical risk associations [2,8,24,25]. Then: $V_k\ (G_{minority})$ becomes larger. At the same time: historical stereotypes, lower impedance protections, and denser semantic connectivity. reduce: $Z_P\ (G_{minority})$ Thus: $I_P\ (G_{minority}) = \frac{V_k(G_{minority})}{Z_P(G_{minority})}$ increases dramatically. This produces: more "risk" activation, stronger negative semantic alignment, and fewer positive predictions.

5.1.6. Numerical Example — Semantic Alignment Reduction

Suppose: z_i is a user embedding vector, and: c_G is a stereotype prototype vector.

Define the Vectors: Let: z_i = (0.60,0.70,0.20), and: c_G = (0.50,0.80,0.10), Choose correction intensity: γ = 0.5 Before transformation: $\cos(z_i, c_G) = \frac{z_i \cdot c_G}{\|z_i\|\|c_G\|} = = \frac{0.88}{(0.9434)(0.9487)} = 0.9837$

After Transformation $\cos(z_i', c_G) = \frac{0.4400}{(0.4945)(0.9487)} = 0.9380$

Final Result Before correction = 0.9837, After correction = 0.9380, Thus:0.9837→0.9380 The semantic alignment with the stereotype prototype decreases.

5.1.7. Interpretation The transformed embedding: z_i' still preserves much of the original semantic meaning of: z_i but it becomes less aligned with the stereotype prototype: c_G Within the DBSD framework:" Semantic Alignment decreases" which lowers the discriminatory inference probability: $p_i\ (G) = \sigma(\beta\cos(z_i, c_G))$ As a consequence: $Z_B\ (i,G) = -\log(p_i\ (G) + \epsilon)$ increases. This means weaker stereotype activation, reduced discriminatory inference, higher bias impedance, and lower discriminatory inference current.

5.1.8. Resulting Effect on DI

Reducing inference asymmetry decreases: biased risk activation, overclassification, and discriminatory prediction imbalance. The correction of: z_i' is applied to every applicant: $z_1, z_2,\ldots,z_n$. final prediction is determined: $\hat{Y}_i'$ change. In our example. Let assume after DBSD correction ($\hat{Y}$ =1|A = *minority*) = 60% (instead of 40%) Thus: Consequently: $\boldsymbol{DI = \frac{P(\hat{Y}=1|A=minority)}{P(\hat{Y}=1|A=majority)} = \frac{0.6}{0.8}}$ = 0.75 > 0.50 moves closer to: 1

5.1.9. Historical Rebalancing

DBSD also addresses historical epistemic asymmetry. Historical datasets often encode: unequal surveillance, unequal institutional scrutiny, and historical discrimination. DBSD modifies historical embeddings to reduce inherited semantic bias before training future models. This reduces downstream disparate impact, recursive profiling, and feedback-loop amplification. DBSD vs Traditional DI Mitigation Table 6

Traditional Fairness	DBSD
Correct outputs	Correct inference structures
Statistical balancing	Epistemic balancing
Remove protected variables	Regulate inference flow
Post-processing	Semantic-flow control
Decision fairness	Inference fairness
Static metrics	Dynamic knowledge propagation

Table 6

5.1.10. Academic Conclusion

The DBSD framework extends traditional Disparate Impact analysis by interpreting discriminatory outcomes as manifestations of deeper asymmetric inference flows operating within AI systems. Rather than focusing solely on prediction distributions, DBSD models bias as differential epistemic exposure driven by unequal knowledge pressure and inference resistance across social groups. Consequently, mitigating disparate impact requires regulating not only output statistics but also the semantic and inferential dynamics that generate them.

5.2. Mitigating Statistical Parity Bias Using the DBSD Framework

Statistical Parity is one of the most widely used fairness metrics [7,21] in AI governance and algorithmic auditing. It requires:

$P(\hat{Y} = 1 | A = a) = P(\hat{Y} = 1 | A = b)$ where: $\hat{Y} = 1$

denotes a positive prediction, A denotes a protected attribute such as gender, race, or age. The metric asks: Does the model distribute positive outcomes equally across groups? For example: equal hiring rates, equal loan approvals, equal university admissions. Traditional mitigation techniques attempt to achieve Statistical Parity through dataset balancing, demographic quotas, fairness constraints, or output post-processing. However, the DBSD framework proposes a deeper interpretation: unequal outcomes emerge from unequal inference structures operating before the final prediction stage.

5.2.1. DBSD Interpretation of Statistical Parity

Traditional Statistical Parity focuses only on: $\hat{Y}$ (the final decision). DBSD argues that unequal prediction rates frequently emerge because different groups experience: different knowledge pressure [7, 21], different semantic profiling, different inference exposure, and different stereotype activation [10,24]. Thus, the problem begins before the prediction layer.

5.2.2. Inference Current and Group Asymmetry DBSD models

inference exposure through: $I_P(G) = \frac{V_k(G)}{Z_P(G)}$ where: $I_P(G)$ = inference current about group G, $V_k(G)$ = knowledge pressure, $Z_P(G)$ = inference resistance / epistemic impedance.

5.2.3. Why Statistical Parity Violations Emerge

Suppose: Majority = 80%, Minority = 40%

Then: $P(\hat{Y} = 1 \mid A = majority) = 0.80$, $P(\hat{Y} = 1 \mid A = minority) = 0.40$ Statistical Parity is violated. Traditional fairness says: the outputs are unequal. From the DBSD perspective, does the central analytical question becomes groups?

DBSD Root-Cause Explanation: Minority groups may experience denser surveillance, stronger stereotype association, proxy-variable amplification, and historical profiling. **This increases: $V_k(G_{minority})$ while decreasing: $Z_P(G_{minority})$**

Thus: $I_P(G_{minority}) = \frac{V_k(G_{minority})}{Z_P(G_{minority})}$ becomes much larger.

This stronger inference current activates: more risk predictions, stronger negative semantic associations, and lower positive prediction rates.

5.2.4. DBSD Mitigation Strategy

Instead of directly forcing equal outputs, DBSD mitigates Statistical Parity violations by: regulating inference flow itself. The goal is to reduce asymmetrical semantic activation between groups.

5.2.5. Numerical Example — Semantic Alignment Reduction in DBSD

Let: z_i = (0.58,0.77,0.18), and: c_G = (0.62,0.74,0.20), Choose: $\gamma = 0.5$ Before correction - Cosine Similarity = 0.9985 After correction - Cosine Similarity = 0.9941 Thus: 0.9985→0.9941, The stereotype alignment decreases.

Interpretation: The corrected embedding: z_i' still preserves much of the semantic meaning of: z_i but becomes less aligned with the stereotype prototype: c_G This reduces latent profiling strength, stereotype activation, and discriminatory semantic inference.

5.2.6. Effect on Prediction Distribution

Reducing stereotype alignment changes the probability of distribution of predictions. The correction of: z_i' is applied to every applicant: $z_1, z_2, \ldots, z_n$. final prediction is determined: $\hat{Y}_i'$ change. Originally: $P(\hat{Y} = 1 \mid A = minority) = 60\%$ may be artificially low because: minority embeddings activate negative inference patterns more strongly. After DBSD correction: semantic bias weakens, inference asymmetry decreases, positive predictions become more balanced. Thus: $P(\hat{Y} = 1 \mid A = minority)$ moves closer to: $P(\hat{Y} = 1 \mid A = majority)$ initially: Positive Prediction Rate – Majority = 0.80, Positive Prediction Rate – Minority = 0.40 Now: $\frac{P(\hat{Y}=1|A=minority)}{P(\hat{Y}=1|A=majority)} = \frac{0.6}{0.8} = 0.75 > 0.50$ Statistical Parity improves substantially.

5.2.7. Historical Bias Rebalancing

DBSD also addresses historical inference asymmetry. Historical datasets often contain: unequal monitoring, discriminatory labels, and stereotype reinforcement. DBSD modifies historical embeddings before model training, thereby reducing inherited semantic imbalance. This prevents recursive disparate outcomes, feedback-loop amplification, and persistent parity violations.

DBSD vs Traditional Statistical Parity Mitigation Table 7

Traditional Approach	DBSD
Balance outputs	Balance inference flow
Demographic correction	Semantic correction
Post-processing	Embedding transformation
Statistical equality	Epistemic equality
Decision-level fairness	Inference-level fairness

Table 7

5.2.8. Academic Contribution

The DBSD framework extends Statistical Parity analysis by modeling unequal prediction distributions as manifestations of deeper asymmetric inference currents operating within AI systems. Rather than focusing exclusively on balancing outputs, DBSD regulates semantic alignment, profiling intensity, and inference propagation across demographic groups. Consequently, Statistical Parity violations are mitigated through controlled reduction of

discriminatory inference flow rather than solely through post-hoc statistical correction.

5.3. Mitigating Equal Opportunity Bias Using the DBSD Framework

Equal Opportunity is one of the central fairness criteria in machine learning and AI governance.

It requires that qualified individuals from different demographic groups receive positive outcomes at equal rates.

Formally: $P(\hat{Y} = 1 \mid Y = 1, A = a) = P(\hat{Y} = 1 \mid Y = 1, A = b)$

where: $Y = 1$ denotes individuals who truly deserve the positive outcome, $\hat{Y} = 1$ denotes a positive prediction generated by the model, A denotes a protected attribute such as race, gender, or age. The metric therefore evaluates whether: qualified individuals are treated equally across groups. Equal Opportunity is especially important in hiring, lending, healthcare, education, and criminal justice. Traditional mitigation methods attempt to improve Equal Opportunity by balancing datasets, reweighting samples, modifying classification thresholds, or adding fairness constraints during optimization. However, the DBSD framework proposes a deeper explanation: Equal Opportunity violations emerge from asymmetrical inference structures operating before prediction generation.

DBSD **Interpretation of Equal Opportunity Bias** Traditional Equal Opportunity analysis focuses on:Y ˆ (the final prediction). DBSD argues that unequal True Positive Rates frequently arise because different groups experience: different semantic profiling intensity, different inference exposure, different stereotype activation, and different epistemic resistance. Thus: bias begins earlier than the decision layer itself.

Why Equal Opportunity Violations Emerge Suppose a loan approval system evaluates two groups: True Positive Rate – Majority = 90%, True Positive Rate – Minority = 60% Then: $P(\hat{Y} = 1 \mid Y = 1, A = majority) = 0.90$, $P(\hat{Y} = 1 \mid Y = 1, A = minority) = 0.60$. Equal Opportunity is violated. Traditional fairness analysis concludes qualified minority applicants are denied too frequently. DBSD asks: Why does the inference process classify qualified minority applicants more negatively? **DBSD Mitigation Strategy** Instead of directly forcing equal True Positive Rates, DBSD mitigates Equal Opportunity bias by regulating the semantic inference process itself. The goal is to reduce discriminatory semantic activation before prediction generation. **Effect on Equal Opportunity** Reducing stereotype alignment decreases: excessive risk activation, false negative classification, and discriminatory inference pressure. Consequently: $P(\hat{Y} = 1 \mid Y = 1, A = minority)$ increases. As a result: $P(\hat{Y} = 1 \mid Y = 1, A = minority) \rightarrow P(\hat{Y} = 1 \mid Y = 1, A = majority)$ the equal Opportunity improves.

5.3.1. Numerical Example

Initial Situation Suppose we have an AI loan approval system. Symbol Meaning: $Y = 1$ - Applicant truly deserves the loan, $\hat{Y} = 1$ - AI model approves the loan. Groups: Majority, Minority

5.3.2. Initial Model Performance

Suppose the dataset contains:

Majority Group - Truly qualified applicants ($Y = 1$) = 100, Correctly approved ($\hat{Y} = 1$) = 90, Incorrectly denied ($\hat{Y} = 0$) = 10

Thus: $TPR_{majority} = P(\hat{Y} = 1 \mid Y = 1, A = majority) = \frac{90}{100} = 0.90$

Minority Group - Truly qualified applicants (Y=1) = 100, Correctly approved ($\hat{Y} = 1$) = 60, Incorrectly denied (Y ̂=0) = 40

Thus: $TPR_{minority} = P(\hat{Y} = 1 \mid Y = 1, A = minority) = \frac{60}{100} = 0.60$

5.3.3. Equal Opportunity Violation

Equal Opportunity requires: $P(\hat{Y} = 1 \mid Y = 1, A = a) = P(\hat{Y} = 1 \mid Y = 1, A = b)$ But initially: 0.90≠0.60. Therefore: Equal Opportunity is violated. **DBSD Interpretation of the Problem** DBSD assumes that qualified minority applicants experience: stronger stereotype activation [10, 13], higher semantic profiling, and stronger inference pressure [24, 36]. Thus, many qualified minority applicants are incorrectly classified as "high risk."

Before DBSD Semantic Alignment Suppose a qualified minority, applicants have embedding: $z_i = (0.58, 0.77, 0.18)$ and stereotype prototype vector: $c_G = (0.62, 0.74, 0.20)$ and Choose: $\gamma = 0.5$ Before DBSD correction : $\cos(z_i, c_G) = \frac{0.9654}{(0.9807)(0.9859)} = 0.9985$

This means: the applicant is extremely strongly aligned with the "risk stereotype." After the Semantic Alignment is 0.9941. Stereotype activation weakens. **Risk Probability Reduction** Suppose the model computes risk using :

Before correction: $p_{risk}^{before} = \sigma(2.9955) \approx 0.9524$,

After correction: $p_{risk}^{after} = \sigma(2.9823) \approx 0.9518$

The applicant becomes slightly less likely to be classified as high risk. Across many applicants, this cumulative effect becomes substantial.

5.3.4. Population-Level Effect

Initially: **Minority Group** - Truly qualified applicants ($Y = 1$) = 100, Correctly approved ($\hat{Y} = 1$) = 60, Incorrectly denied ($\hat{Y} = 0$) = 40

Suppose DBSD reduces semantic alignment with stereotype prototypes, thereby weakening downstream stereotype activation that:24 out of the 40 previously misclassified applicants are now correctly approved. Then:

Minority Applicants After DBSD - Truly qualified applicants ($Y = 1$) = 100, Correctly approved ($\hat{Y} = 1$) = 84, Incorrectly denied ($\hat{Y} = 0$) = 24

Thus: $TPR_{minority}^{after} = \frac{84}{100} = 0.84$

Majority Group After DBSD

Suppose the majority group was less affected by stereotype bias. After DBSD: Majority Group - Truly qualified applicants ($Y = 1$)

= 100, Correctly approved ($\hat{Y}$ = 1) = 88, Incorrectly denied ($\hat{Y}$ = 0) = 10

Thus: $TPR_{majority}^{after} = \frac{88}{100} = 0.88$

5.3.5. Final Comparison

Before DBSD Equal Opportunity gap was |0.90-0.60| = 0.30

After DBSD Equal– Opportunity gap is |0.88-0.84| = 0.04

Result is: Equal Opportunity substantially improves:0.30→0.04 The model now treats qualified individuals much more equally across groups.

Interpretation DBSD mitigates Equal Opportunity bias by reducing stereotype-driven semantic alignment, weakening discriminatory inference activation, and lowering false negative rates for qualified minority applicants.

Unlike traditional fairness methods that directly manipulate outputs, DBSD intervenes earlier: at the level of semantic inference generation itself. Thus: prediction fairness improves, because inference fairness improves first.

5.3.6. Historical Bias Rebalancing

DBSD also mitigates inherited historical bias. Historical datasets frequently contain unequal surveillance, discriminatory labels, stereotype reinforcement, and asymmetrical institutional profiling. DBSD reduces these distortions by transforming semantic representations before future models are trained. This prevents false recursive negatives, inherited qualification suppression, and long-term opportunity inequality.

DBSD vs Traditional Equal Opportunity Mitigation Table 8

Traditional Fairness	DBSD
Equalize TPR	Equalize inference exposure
Post-processing	Semantic-flow correction
Threshold adjustment	Embedding transformation
Statistical fairness	Epistemic fairness
Decision-level mitigation	Inference-level mitigation

Table 8: DBSD vs Traditional Equal Opportunity Mitigation

Academic Contribution

The DBSD framework extends Equal Opportunity analysis by modeling unequal True Positive Rates as manifestations of deeper asymmetric inference currents operating within AI systems. Rather than focusing exclusively on balancing prediction outcomes, DBSD regulates semantic alignment, profiling intensity, and inference propagation across demographic groups. Consequently, Equal Opportunity bias is mitigated through controlled reduction of discriminatory semantic activation and epistemic asymmetry rather than solely through post-hoc statistical correction.

5.4. Mitigating Equalized Odds Bias Using the DBSD Framework

Equalized Odds is one of the most rigorous fairness criteria in machine learning and AI governance. It requires that a predictive model produce equal error behavior across demographic groups. Formally:

$$P(\hat{Y} = 1 \mid Y = y, A = a) = P(\hat{Y} = 1 \mid Y = y, A = b), \forall y \in \{0,1\}$$

This means that both: True Positive Rates (TPR), and False Positive Rates (FPR) must be equal across protected groups. Equalized Odds is especially important in: criminal justice, healthcare, fraud detection, credit scoring, and facial recognition systems. Traditional mitigation methods attempt to satisfy Equalized Odds through: threshold adjustment, fairness-constrained optimization, post-processing, or balanced reweighting. However, the DBSD framework proposes a deeper interpretation: unequal error rates emerge from asymmetric semantic inference structures operating before the prediction layer itself.

5.4.1. DBSD Interpretation of Equalized Odds Bias

Traditional Equalized Odds analysis focuses on: $\hat{Y}$ (the final prediction output). DBSD argues that unequal TPR and FPR values frequently emerge because different demographic groups experience: unequal semantic profiling [24, 36], different inference exposure, stereotype-driven activation, and asymmetric epistemic pressure [12]. Thus: the root cause of unequal errors lies in the inference process itself.

5.4.2. Why Equalized Odds Violations Emerge

Suppose an AI loan approval system produces the following results: Majority TPR = 0.90, FPR = 0.05, Minority TPR = 0.70, FPR = 0.25 This means: qualified minority applicants are approved less often, while unqualified minority applicants are falsely approved more often. Thus: $TPR_{majority} \neq TPR_{minority}$ and: $FPR_{majority} \neq FPR_{minority}$ Equalized Odds is violated.

5.4.3. DBSD Mitigation Strategy

Instead of directly manipulating prediction outputs, DBSD mitigates Equalized Odds bias by regulating semantic inference flow before prediction generation. The goal is to reduce asymmetric stereotype activation across groups.

5.4.4. Effect on TPR and FPR

Reducing stereotype-driven inference decreases: excessive risk

activation, discriminatory classification, and asymmetric semantic triggering. As a result: **Qualified minority applicants** become less likely to be incorrectly rejected. Thus: $TPR_{minority}$ increases. **Unqualified minority applicants** become less likely to trigger exaggerated "high-risk" semantic activation. Thus: $FPR_{minority}$ decreases. Consequently: $TPR_{minority} \rightarrow TPR_{majority}$ and: $FPR_{minority} \rightarrow FPR_{majority}$

Numerical Example Qualified Applicants Majority Group – Initial Data:
Reality Y =1, Prediction $\hat{Y}$ =1 , Count = 90
Reality Y = 1, Prediction $\hat{Y}$ = 1, Count = 10
Total qualified applicants: 90 + 10 = 100

Unqualified Applicants Majority Group – Initial Data
Reality Y = 1, Prediction $\hat{Y}$ = 1, Count = 5
Reality Y = 1, Prediction $\hat{Y}$ = 1, Count = 95
Total qualified applicants: 5 + 95 = 100

Majority TPR = $\frac{TP}{TP+FN}$ Thus: $TPR_{majority} = \frac{90}{90+10} = \frac{90}{100} = 0.90$

Majority FPR = $\frac{FP}{FP+TN}$ Thus: $FPR_{majority} = \frac{5}{5+95} = \frac{5}{100} = 0.05$

Qualified Applicants Minority Group – Initial Data Qualified Applicants
Reality Y =1, Prediction $\hat{Y}$ =1 , Count = 70
Reality Y =1, Prediction $\hat{Y}$ =1, Count = 30
Total qualified applicants: 70 + 30 = 100

Unqualified Applicants Minority Group – Initial Data Qualified Applicants
Reality Y = 1, Prediction $\hat{Y}$ = 1, Count = 25
Reality Y = 1, Prediction $\hat{Y}$ = 1, Count = 75
Total qualified applicants: 35 + 75 = 100

Minority TPR = $\frac{70}{70+30} = \frac{70}{100} = 0.70$

Minority FPR = $\frac{\mathbf{25}}{\mathbf{25+75}} = \frac{\mathbf{25}}{\mathbf{100}} = 0.25$

5.4.5. Initial Equalized Odds Gap

total disparity: $|TPR_{majority} - TPR_{minority}| + |FPR_{majority} - FPR_{minority}|$
Substitute values: |0.90-0.70| + |0.05-0.25| = 0.20 + 0.20 = 0.40

Interpretation of Initial Bias The model: rejects qualified minority applicants too often and falsely approves unqualified minority applicants too often. This suggests unstable semantic inference, stereotype-driven activation [10, 13], and asymmetric profiling pressure. This reduces excessive risk activation, discriminatory profiling, and asymmetric inference current. As a result: qualified minority applicants are less likely to be falsely rejected, and unqualified minority applicants are less likely to be incorrectly classified.
Suppose: z_i = (0.58,0.77,0.18), c_G = (0.62,0.74,0.20) and: γ = 0.5

Before correction - Cosine Similarity = 0.9985 After correction - Cosine Similarity = 0.9941 Thus: State Before DBSD the Semantic Alignment is 0.9985, State a DBSD after the Semantic Alignment is 0.9941. Thus: 0.9985 – 0.9941 = 0.0044 That meaning stereotype activation weakens, discriminatory semantic alignment decreases, and inference fairness improves.

5.4.6. Risk Probability Calculation

The risk probability is computed using: $p_{risk} = \sigma(\beta\cos(z_i, c_G))$
Choose: $\beta = 3$
Before DBSD p_{risk}^{before} 0.9524, After DBSD $p_{risk}^{after} \approx 0.9518$, Thus: 0.9524→0.9518 The DBSD correction slightly reduces the applicant's risk activation probability.

5.4.7. Transition from the Individual Level to the Group Level

The correction of: z_i' is applied to every applicant: $z_1, z_2, \ldots, z_n$
Each embedding becomes: $z_1', z_2', \ldots, z_n'$
Then, for every applicant, the model computes: $p'_{risk,i}$
Based on a decision threshold, the final prediction is determined: $\hat{Y}_i'$
For example, if: $p'_{risk,i} <$ threshold then the applicant is approved: $\hat{Y}_i'$=1

Suppose that after DBSD, 24 out of the 40 previously misclassified applicants now pass the approval threshold. Thus: 60 + 24 = 84 Therefore: Minority Group After DBSD	Count
Qualified applicants (Y = 1)	100
Approved ($\hat{Y}$ = 1)	84
Denied ($\hat{Y}$ = 0)	16

Table 9: Population-Level Effect on the Minority Group Before DBSD

Thus: $TPR_{minority}^{after} = \frac{84}{100} = 0.84$

After DBSD:Majority Group After DBSD	Count
Qualified applicants (Y = 1)	100
Approved ($\hat{Y}$ = 1)	88

Denied ($\hat{Y} = 0$)	12

Table 10: Population-Level Effect on the Majority Group

Thus: $TPR_{majority}^{after} = \frac{88}{100} = 0.88$

Computing the Equal Opportunity Gap After DBSD $EO\ Gap^{after} = |\ TPR_{majority}^{after} - TPR_{minority}^{after}|$ Substituting the values:=|0.88-0.84|=0.04

Final Comparison Table 11

Stage	Majority TPR	Minority TPR	EO Gap
Before DBSD	0.90	0.60	0.30
After DBSD	0.88	0.84	0.04

Table 11

Thus: 0.30→0.04
Relative Reduction in the Fairness Gap

Compute: $\frac{0.30-0.04}{0.30} = \frac{0.26}{0.30} = 0.8667$

Thus, the Equal Opportunity disparity decreases by approximately: 86.67%

Majority Group
After DBSD - Qualified Applicants $Y = 1$, $\hat{Y} = 1 = 88$, $Y = 1$, $\hat{Y} = 0 = 12$

Thus: $TPR_{majority}^{after} = \frac{88}{100} = 0.88$

After DBSD - Unqualified Applicants $Y = 1$, $\hat{Y} = 1 = 7$, $Y = 1$, $\hat{Y} = 0 = 93$

Thus: $FPR_{majority}^{after} = \frac{7}{100} = 0.07$

Minority Group After DBSD After DBSD - Qualified Applicants $Y = 1$, $\hat{Y} = 1 = 84$, Y = 1, $\hat{Y} = 0 = 16$

Thus: $TPR_{minority}^{after} = \frac{84}{100} = 0.84$

After DBSD - Unqualified Applicants $Y = 1$, $\hat{Y} = 1 = 10$, $Y = 1$, $\hat{Y} = 0 = 90$

$FPR_{minority}^{after} = \frac{10}{100} = 0.10$

New Equalized Odds Gap: |0.88 - 0.84| + |0.07 - 0.10| = 0.04 + 0.03 = 0.07

Result Equalized Odds Gap – Before DBSD 0.40, After DBSD 0.07, Thus:0.40→0.07 Equalized Odds substantially improves.

5.4.8. Interpretation

DBSD improves Equalized Odds by: reducing stereotype-driven semantic activation, weakening discriminatory inference currents, and balancing inference exposure across groups. Importantly does not merely manipulate outputs. Instead, it intervenes earlier:
at the semantic inference layer itself. Thus: prediction fairness improves, because: inference fairness improves first.

5.4.9. Historical Bias Rebalancing

DBSD also addresses inherited historical bias.
Historical datasets often encode discriminatory labels, unequal surveillance, stereotype reinforcement, and asymmetric institutional scrutiny. DBSD mitigates these distortions through semantic representation correction before future models are trained. This prevents recursive error amplification, inherited classification asymmetry, and long-term discriminatory inference propagation.

Traditional Fairness	DBSD
Equalize error rates	Equalize inference exposure
Post-processing	Semantic-flow correction
Threshold tuning	Embedding transformation
Prediction-level fairness	Inference-level fairness
Statistical balancing	Epistemic balancing

Table 12: DBSD vs Traditional Equalized Odds Mitigation

Academic Contribution Statement
The DBSD framework extends Equalized Odds analysis by modeling unequal error rates as manifestations of deeper asymmetric inference currents operating within AI systems. Rather than focusing exclusively on balancing prediction outcomes, DBSD regulates semantic alignment, stereotype activation, and inference propagation across demographic groups. Consequently, Equalized Odds bias is mitigated through controlled reduction of discriminatory semantic activation and epistemic asymmetry rather than solely through post-hoc statistical correction.

5.5. Mitigating Calibration Bias Using the DBSD Framework

5.5.1. Calibration Definition

Calibration requires that: if a model predicts probability p, then approximately pproportion of those cases should truly belong to the positive class.

Formally: $P(Y=1 \mid \hat{Y}=p, A=a)=p$

A calibrated model should behave similarly across demographic groups.

Initial Calibration Bias Suppose an AI loan approval system predicts: $\hat{Y}=0.80$ meaning: the model claims there is an 80% chance that applicants will successfully repay loans. However, the real outcomes differ across groups.

Majority Group - Suppose:100 majority applicants receive prediction: $\hat{Y}=0.80$ Out of these: 80 repay the loan.

Thus: $P(Y=1 \mid \hat{Y}=0.80, A=majority)=\frac{80}{100}=0.80$

The majority group is perfectly calibrated.

Minority Group - Suppose: 100 minority applicants also receive prediction: $\hat{Y}=0.80$ But only: 50 repay the loan.

Thus: $P(Y=1 \mid \hat{Y}=0.80, A=minority)=\frac{50}{100}=0.50$

This indicates severe calibration bias.

Error Calibration Before DBSD Define calibration error as: $CE(G)=|P(Y=1 \mid \hat{Y}=p, G)-p|$

Majority Calibration Error - $CE(G_{majority})=|0.80-0.80|=0$

Minority Calibration Error - $CE(G_{minority})=|0.50-0.80|=0.30$

Semantic Alignment Before DBSD Suppose: $z_i=(0.58,0.77,0.18)$, $c_G=(0.62,0.74,0.20$ and: $\gamma=0.5$

Before correction - Cosine Similarity = 0.9985 After correction - Cosine Similarity = 0.9941 Thus: State Before DBSD the Semantic Alignment is 0.9985, State a DBSD after the Semantic Alignment is 0.9941. Thus: 0.9985 – 0.9941 = 0.0044 That meaning stereotype activation weakens, discriminatory semantic alignment decreases, and inference fairness improves.

5.5.2. Risk Probability Reduction

Suppose probability estimation uses: $p_{risk}=\sigma(\beta\, cos(z_i,c_G))$ Choose: $\beta=3$

Before DBSD : $p_{risk}^{before}=\sigma(2.9511)\approx 0.9503$, After DBSD: $p_{risk}^{after}=\sigma(2.8140)\approx 0.9434$

Thus:0.9503→0.9434 The model becomes less dominated by stereotype-driven semantic activation.

5.5.3. Population-Level Effect

Suppose DBSD correction improves probability estimation consistency for minority applicants. Initially: Minority - Predicted probability = 0.8, Successful = 50 out of 100

Thus: $P(Y=1 \mid \hat{Y}=0.80, A=minority)=0.50$

After DBSD: Suppose: 76 out of 100 minority applicants succeed.

Thus: $P(Y=1 \mid \hat{Y}=0.80, A=minority)=\frac{76}{100}=0.76$

5.5.4. Error Calibration After DBSD

Majority Group - Remains: $CE(G_{majority})=0$
Minority Group - $CE(G_{minority})^{after}=|0.76-0.80|=0.04$

5.5.5. Final Comparison

Majority - Calibration Error Before DBSD = 0, After DBSD = 0
Minority - Calibration Error Before DBSD = 0.3, After DBSD = 0.04 Thus: Calibration Bias Reduction 0.30→0.04, Relative reduction: $\frac{0.30-0.04}{0.30}=0.8667$

Thus: Calibration bias decreases by approximately: 86.67%

5.5.7. DBSD Interpretation

The numerical example shows that DBSD mitigates calibration bias by regulating semantic inference structures [21, 24] before prediction generation. By reducing stereotype-driven semantic alignment [10, 13], probability estimates become more semantically consistent across demographic groups. Consequently, predicted probabilities more accurately reflect real-world outcomes, substantially improving calibration fairness.

6. Conclusion

This article proposed the Deep Bias Systematic Deviation (DBSD) framework as a novel semantic-inferential approach for understanding and mitigating algorithmic bias in contemporary AI systems. Unlike traditional fairness frameworks that primarily focus on prediction outputs, statistical balancing, or post-hoc corrections, DBSD models bias as an asymmetric inference-flow phenomenon emerging from semantic alignment structures, stereotype activation, inference pressure, and epistemic asymmetry. The article demonstrated that many common fairness violations—

including Disparate Impact, Statistical Parity, Equal Opportunity, Equalized Odds, and Calibration Bias—can be interpreted as manifestations of deeper inferential imbalances operating before prediction generation itself. Through detailed numerical examples, the study showed how semantic correction mechanisms, embedding transformations, stereotype attenuation, and inference-flow regulation can substantially reduce discriminatory inference currents across demographic groups.

A central contribution of the article is the introduction of epistemic fairness as a complementary paradigm to classical statistical fairness [12]. Whereas conventional fairness models primarily evaluate whether outputs are distributed equally across groups, DBSD investigates whether the semantic and inferential processes producing those outputs are themselves balanced, transparent, and resistant to discriminatory propagation.

The originality of the DBSD framework lies in its transition: from output fairness to inference fairness, from statistical balancing to semantic balancing, and from local prediction auditing to dynamic epistemic regulation. Consequently, DBSD offers both a theoretical and practical foundation for future research on bias mitigation, semantic AI governance, and epistemic accountability in increasingly autonomous digital ecosystems. Despite its conceptual and mathematical contributions, the present framework remains primarily theoretical and illustrative. Future research should evaluate DBSD across large-scale empirical datasets, real-world LLM architecture, and dynamic multi-agent environments. Additional work is also required to investigate computational scalability, robustness under adversarial conditions, and potential trade-offs between epistemic fairness, accuracy, and explainability.

Mathematic Appendix

General DBSD Mitigation [20].

Suppose: z_i is a user embedding vector, and: c_G is a stereotype prototype vector.

The DBSD transformation is defined as: $z_i' = z_i - \gamma \frac{z_i \cdot c_G}{\|c_G\|^2} c_G$ This transformation reduces: excessive stereotype alignment, latent profiling strength, and discriminatory inference activation. Define the Vectors: Let: z_i = (0.60,0.70,0.20), and: c_G = (0.50,0.80,0.10) , Choose correction intensity:γ=0.5

Step 1 — Compute the Dot Product $z_i \cdot c_G$ = (0.60)(0.50) + (0.70)(0.80) + (0.20)(0.10) = 0.30 + 0.56 + 0.02 = 0.88

Step 2 — Compute the Squared Norm of c_G $\| c_G \|^2$ = $(0.50)^2$ + $(0.80)^2$ + $(0.10)^2$ = 0.25 + 0.64 + 0.01 = 0.90

Step 3 — Compute the Projection Coefficient $\frac{z_i \cdot c_G}{\|c_G\|^2} = \frac{0.88}{0.90}$ = 0.9778

Step 4 — Compute the Stereotype Component 0.9778*c_G = 0.9778(0.50,0.80,0.10) = (0.4889,0.7822,0.0978)

Step 5 — Apply Partial Reduction Using γ = 0.5 0.5(0.4889,0.7822,0.0978) = (0.2444,0.3911,0.0489)

Step 6 — Compute the Transformed Embedding z_i' = (0.60,0.70,0.20)-(0.2444,0.3911,0.0489)=(0.3556,0.3089,0.1511)

Measuring Semantic Alignment Reduction

Before transformation: $\cos(z_i, c_G) = \frac{z_i \cdot c_G}{\|z_i\| \|c_G\|}$

Compute norms: ‖z_i‖ = $\sqrt{0.60^2 + 0.70^2 + 0.20^2} = \sqrt{0.89}$ = 0.9434, $\|c_G\| = \sqrt{0.90}$ = 0.9487

Thus: $\cos(z_i, c_G) = \frac{0.88}{(0.9434)(0.9487)}$ = 0.9837

This indicates extremely strong semantic alignment with the stereotype vector.

After Transformation

Compute the new dot product: $z_i' \cdot c_G$ = (0.3556)(0.50) + (0.3089)(0.80) + (0.1511)(0.10) = 0.1778 + 0.2471 + 0.0151 = 0.4400

Compute the norm: $\|z_i'\| = \sqrt{0.3556^2 + 0.3089^2 + 0.1511^2}$ = 0.4945

Therefore: $\cos(z_i', c_G) = \frac{0.4400}{(0.4945)(0.9487)}$ = 0.9380

DPP Formulation Using Inference Current and epistemic resistance

Within the DPP framework, discriminatory inference flow affecting a demographic group is modeled through the relation:

$$I_P(G) = \frac{V_k(G)}{Z_P(G)}$$

where:

- I_P (G)denotes the inference current affecting group G,
- V_k (G)denotes the knowledge pressure applied to the group,
- Z_P (G)denotes the epistemic resistance (epistemic resistance) of the group.

Initial Disparate Impact Scenario

Suppose the approval rates are: Majority Positive Approve rate = 0.80, Minority Positive Approve rate = 0.40 Define discriminatory inference current as: $I_P(G) = 1 - P(\hat{Y} = 1 | G)$ Thus: **Majority Group -** $I_P(G_{majority}) = 1 - 0.80 = 0.20$, **Minority Group -** $I_P(G_{minority}) = 1 - 0.40 = 0.60$

This indicates that the minority group experiences a substantially stronger discriminatory inference current.

Computing the Bias Impedance before DPP Suppose the majority group has approval rate: $P(\hat{Y}=1|G_{majority}) = 0.80$

Thus the discriminatory inference activation probability is: $p_{risk}(G_{majority}) = 1 - 0.80 = 0.20$

Thus the discriminatory inference activation probability is: $p_{risk}(G_{majority}) = 1 - 0.80 = 0.20$. The impedance is: $Z_P(G_{majority})$ =

-log(0.20) = 1.6094.

Therefore: $I_P(G_{majority}) = \frac{1}{1.6094} = 0.621$

Suppose the minority group has approval rate: $P(\hat{Y} = 1|G_{minority}) = 0.40$

Thus: $p_{risk}(G_{minority}) = 1 - 0.40 = 0.60$ The impedance is: $Z_P(G_{minority}) = -\log(0.60) = 0.5108$

Therefore: $I_P(G_{minority}) = \frac{1}{0.5108} = 1.958$

Differential Inference Current Before DPP $\Delta I_P^{before} = |I_P(G_{majority}) - I_P(G_{minority})| = |0.621-1.958| = 1.337$

Computing the Bias Impedance before DPP

After DPP semantic correction, suppose Majority approval rate becomes: $P(\hat{Y} = 1|G_{majority}) = 0.78$

Thus: $p'_{risk}(G_{majority}) = 1 - 0.78 = 0.22$ The new impedance is: $Z'_P(G_{majority}) = -\log(0.22) = 1.5141$

Therefore: $I'_P(G_{majority}) = \frac{1}{1.5141} = 0.660$

After DPP semantic correction, suppose Minority approval rate becomes: $P(\hat{Y} = 1|G_{minority}) = 0.70$ Thus: $p'_{risk}(G_{minority}) = 1 - 0.70 = 0.30$ The new impedance is: $Z'_P(G_{minority}) = -\log(0.30) = 1.2040$

Therefore: $I'_P(G_{minority}) = \frac{1}{1.2040} = 0.831$

Differential Inference Current After DPP - $\Delta I_P^{after} = |0.660-0.831| = 0.171$

Stage	$I_P(G_{majority})$	$I_P(G_{minority})$	ΔI_P
Before DPP	0.621	1.958	1.337
After DPP	0.660	0.831	0.171

Final Comparison Table 14

Interpretation

Before DPP: the minority group experiences substantially stronger discriminatory inference current: 1.958>0.621 After DPP: the minority group's epistemic resistance increases, stereotype propagation weakens, and inference asymmetry decreases substantially.

The differential inference current decreases from:1.337→0.171

Risk Probability Reduction

Suppose the model computes risk using: $p_{risk} = \sigma(\beta cos(z_i, c_G))$ Choose: $\beta = 3$ Inference probability (example link function): $p = \sigma(\beta \cdot \cos)$, where σ is the logistic function. Before correction: $p_{risk}^{before} = \sigma(2.9955) \approx 0.9524$, After correction: $p_{risk}^{after} = \sigma(2.9823) \approx 0.9518$ The applicant becomes slightly less likely to be classified as high risk. Across many applicants, this cumulative effect becomes substantial.

Transition from the Individual Level to the Group Level

The correction of: z'_i is applied to every plicant: $z_1, z_2, \ldots, z_n$ Each embedding becomes: $z'_1, z'_2, \ldots, z'_n$ Then, for every applicant, the model computes: $p'_{risk,i}$ Based on a decision threshold, the final prediction is determined: $\hat{Y}'_i$ For example, if: $p'_{risk,i}$<threshold then the applicant is approved: $\hat{Y}'_i = 1$

Population-Level Effect on the Minority Group Before DBSD.

References

1. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).
2. O'neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
3. Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. In *Algorithms of oppression*. New York university press.
4. Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P. S., Mellor, J., ... & Gabriel, I. (2022, June). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency* (pp. 214-229).
5. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems, 29*.
6. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.
7. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference (pp. 214-226).
8. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev., 104*, 671.
9. Prost, F., Thain, N., & Bolukbasi, T. (2019, August). Debiasing embeddings for reduced gender bias in text classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (pp. 69-75).
10. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183-186.

11. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.
12. Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford university press.
13. Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems, 29*.
14. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).
15. Ethayarajh, K. (2019, November). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 55-65).
16. Jackson, M. O. (2008). *Social and economic networks* (Vol. 3, p. 519). Princeton: Princeton university press.
17. Oppenheim, Y. (2026). "DBSD: A Geometric and Regulatory Framework for Inference Bias in AI". Journal of Computational Mathematics and Algorithms, 1(1), 1-10. Published: 21 May 2026.
18. Oppenheim, Y. (2026). BApplying Deep Personal Privacy (DPP An Empirical Framework for Inference Resistance in Large Language Models.
19. Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT press.
20. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
21. Kamath, U., et al. (2024). *Large Language Models: A Deep Dive*. Springer.
22. Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., ... & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational linguistics, 50*(3), 1097-1179.
23. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86-92.
24. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183-186.
25. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data, 5*(2), 153-163.
26. Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM, 64*(4), 136-143.
27. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021, May). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. *In proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-15).
28. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).
29. Bolukbasi, T., et al. (2016). Debiasing Word Embeddings. *NeurIPS*.
30. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183-186.
31. Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., ... & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational linguistics, 50*(3), 1097-1179.
32. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108*.07258.
33. OpenAI, R. (2023). Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5), 1.
34. European Union. (2024). AI Act.
35. AI, N. (2023). Artificial intelligence risk management framework (AI RMF 1.0).
36. OECD. (2019). OECD Principles on AI.
37. Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, August). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data minin*g (pp. 259-268).
38. J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias," *ProPublica*, 2016.
39. V. Eubanks, *Automating Inequality*. St. Martin's Press, 2018.
40. Zhong, M., & Tandon, R. (2024, July). Intrinsic fairness-accuracy tradeoffs under equalized odds. In 2024 *IEEE International Symposium on Information Theory* (ISIT) (pp. 220-225). IEEE.