Wissen PUBLICATION GROUP

Journal of Computational Mathematics and Algorithms

Research Article

Volume: 1 | Isssue: 1 | May 2026 | Pages 01–10

DBSD: A Geometric and Regulatory Framework for Inference Bias in AI

Yair Oppenheim

The Lester and Sally Antin Faculty of Humanities, Tel Aviv University, Tel Aviv, Israel

*Corresponding Author: Yair Oppenheim, The Lester and Sally Antin Faculty of Humanities, Tel Aviv University, Tel Aviv, Israel.

Submitted: 08 May 2026 **Accepted:** 14 May 2026 **Published:** 21 May 2026

Citation: Oppenheim, Y. (2026). DBSD: A Geometric and Regulatory Framework for Inference Bias in AI. Journal of Computational Mathematics and Algorithms, 1(1), 1-10. DOI: 10.67927/JComputMathAlgor/2026(1)101.

Abstract

This article introduces the Deep Bias–Systematic Deviation (DBSD) framework, a novel theoretical, mathematical, and regulatory approach to understanding bias in contemporary artificial intelligence systems. Unlike conventional approaches that define bias through unequal outputs or statistical disparities, DBSD reconceptualizes bias as the degree of ease with which systematic deviation may be generated, inferred, and propagated through latent representational structures. The framework advances three primary contributions. First, it proposes an impedance-based model of bias, in which harmful inference corresponds to low resistance within information systems. Second, it develops the Tri-World Embedding Bias Model, representing the observed, true, and ideal worlds as vectors within embedding space and quantifying their relationships using cosine similarity. Third, it introduces a dynamic regulatory model in which intervention strength adapts continuously through feedback mechanisms governed by parameters such as λ and γ. By shifting analysis from output bias to inference bias, DBSD integrates philosophy, machine learning, vector geometry, and regulatory theory into a unified framework. It provides a new foundation for measuring, mitigating, and governing bias in large language models and other AI systems.

Keywords: Deep Bias Systematic Deviation (DBSD), Inference Bias, Algorithmic Bias, Large Language Models (LLMs), AI Ethics, Fairness in Machine Learning, Embedding Spaces, Cosine Similarity, Bias Measurement, Bias Mitigation, Dynamic Regulation, AI Governance, Inferential Privacy, Responsible AI, Feedback-Control Regulation.

Introduction

Bias is commonly defined as a tendency or prejudice toward or against a particular entity, often resulting in unfair outcomes [1,2]. In statistical contexts, bias refers to a systematic deviation from the true value of an estimate [3]. In psychology, bias denotes a cognitive predisposition that influences judgment, frequently in an implicit or unconscious manner [4].

Traditional legal and privacy governance models often focus on explicit variables, visible discrimination, formal rules, or direct data usage, rather than latent inferential [5,6].

While these definitions capture bias as deviation, prejudice, or cognitive tendency, they do not fully account for the structural conditions that enable asymmetric inference—an increasingly important feature of modern AI systems [7,8]. In contemporary machine learning environments, bias may arise not only through explicit discriminatory rules or visible unequal outcomes, but also through hidden inferential mechanisms that transform weak signals into socially consequential decisions [7,9]. 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. However, in the age of artificial intelligence and large language models, systematic bias can emerge even when no explicit discriminatory rule is encoded [10-12].

DBSD addresses this shift by modeling bias as **epistemic impedance**—the degree of resistance that prevents systems from transforming available signals into harmful asymmetric judgments about individuals or groups [13,14]. 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 [9,15,16].

By focusing on the dynamics of knowledge extraction rather than merely visible decisions, the DBSD framework provides a new theoretical and practical foundation for bias engineering, fairness analysis, and regulatory governance in AI-driven environments [7,8]. More broadly, the article argues that future fairness analysis must move beyond output auditing toward **inference auditing,** where the central concern is not only what systems decide, but what they are structurally capable of inferring [7,17].

The Present Article Makes Four Primary Contributions

- First, it introduces the DBSD framework, which reconceptualizes bias as a structural susceptibility to inference rather than a static property of outputs.
- Second, it develops the Tri-World Embedding Bias Model, enabling geometric comparison between the observed, true, and ideal worlds within embedding space.
- Third, it proposes quantitative tools for bias measurement and mitigation based on cosine similarity, impedance, and controlled vector steering.
- Fourth, it advances a dynamic regulatory model in which intervention strength is governed by adaptive feedback mechanisms through parameters such as λ and γ.

Related Work

The contemporary literature on AI bias and fairness has developed along several major lines [7,8]. The first strand focuses on discriminatory outcomes in automated decision systems [9]. Foundational work by Barocas and Selbst demonstrated how data-driven systems may reproduce legally significant disparate impacts even without explicit discriminatory intent. Later work by Barocas, Hardt, and Narayanan further systematized fairness problems in machine learning, emphasizing that different fairness metrics often involve normative and statistical tradeoffs [7,9]. Kleinberg, Mullainathan, and Raghavan further demonstrated that some fairness criteria cannot be simultaneously satisfied, thereby highlighting inherent tradeoffs in algorithmic risk scoring [17].

A second strand addresses the theoretical limits of fairness itself [8,17]. Friedler argued that fairness depends on assumptions regarding the relationship between the observed world, the true world, and the ideal world, implying that no single fairness metric can satisfy all normative goals simultaneously [18]. Related work by Hardt, Price, and Srebro introduced equality of opportunity as a formal fairness criterion centered on error-rate parity [8].

A third strand concerns the rise of large language models and generative AI [19]. Bender warned that increasingly large language models may reproduce social harm embedded in web-scale corpora while obscuring accountability [10]. Weidinger proposed a taxonomy of risks posed by language models, including representational harm, misinformation, and discriminatory outputs [20]. More recently, Gallegos surveyed bias and fairness challenges in LLMs across gender, race, culture, and political domains [12].

Despite these important advances, most existing approaches focus primarily on outputs, predictions, or observable disparities [7,8,18]. The DBSD framework complements this literature by shifting analytical attention toward the **latent inferential pathways** through which asymmetric knowledge is generated before explicit decisions occur. In this sense, DBSD extends fairness research from output bias to **structural inference bias**, thereby offering a new bridge between machine learning, ethics, and regulatory design [7,17].

Bias in the Internet Age: Artificial Intelligence and Bias

Bias in artificial intelligence has emerged as one of the defining ethical, legal, and technical challenges of the digital era [7,21]. As AI systems increasingly mediate access to employment, education, healthcare, finance, security, and public information, growing concern has arisen regarding their capacity to reproduce, amplify, or conceal pre-existing social inequalities. Unlike isolated human decisions, algorithmic systems may operate continuously, at scale, and across millions of users, thereby transforming localized prejudice into systemic outcomes [9,19,21]. Consequently, bias in AI is no longer merely a statistical or engineering problem; it has become a central issue of governance, legitimacy, and social justice [7,9,22].

In broad terms, bias may be understood as a systematic distortion in representation, measurement, judgment, or decision-making that leads to unfair, exclusionary, or inaccurate outcomes. Within AI systems, such distortions may arise at multiple stages of the computational pipeline from data collection and annotation to model optimization, deployment, and post-deployment interaction [23]. Recent scholarship therefore emphasizes that bias is not a single defect located in one model component, but rather a multi-layered phenomenon emerging from the interaction among datasets, optimization objectives, institutional assumptions, and user behavior [18,23].

Main Sources of Bias in AI

A widely accepted analytical distinction identifies several major sources of bias in contemporary AI systems.

Data Bias

Training data may underrepresent particular populations or encode discriminatory historical patterns [23,24]. If past hiring decisions favored one demographic group, predictive hiring systems trained on those records may inherit and reproduce such disparities [9]. Likewise, internet-scale corpora may contain stereotypes, toxic language, or culturally skewed narratives that later shape model behavior [10,12,16].

Label Bias

Many machine learning systems rely on labels produced by human annotators or institutional processes [24,25]. These labels may themselves reflect prejudice, inconsistent norms, or unequal treatment [18,23]. Categories such as “risk,” “quality,” or “trustworthiness” are often socially mediated rather than purely objective, thereby importing normative distortions into technical systems [17,18].

Algorithmic Bias

Even when training data appear relatively balanced, optimization procedures may generate unequal outcomes [7,8]. Models designed to maximize aggregate accuracy often perform best on majority populations while producing higher error rates for minority groups [8,14,15,18]. This issue is especially serious where false positives or false negatives impose asymmetric harms [8].

Interaction Bias

AI systems that interact with users may absorb, reinforce, or adapt to harmful behavioral patterns [12,20]. Feedback loops may emerge when users reward sensationalism, extremity, or discriminatory outputs [12]. Conversational systems are particularly vulnerable to such dynamic forms of bias [12].

Deployment Bias

A model may perform adequately in one context yet become unfair when transferred to another environment [23,26]. Systems trained on one linguistic, cultural, or socioeconomic population

may produce distorted outcomes when deployed elsewhere [12,16,23,24]. Bias therefore often arises not only from model design, but from inappropriate real-world application [23].

Bias in Large Language Models (LLMs)

The rapid development of large language models has intensified concerns regarding bias [11,19]. Unlike narrower predictive systems, LLMs generate language, summarize information, rank alternatives, translate text, and increasingly function as interfaces between humans and knowledge systems [11]. Because they are trained on vast internet corpora, they inevitably absorb many of the asymmetries, exclusions, conflicts, and stereotypes embedded in digital culture [10,16].

Recent studies indicate that LLMs may express or reproduce stereotypes relating to gender, race, religion, nationality, age, disability, and political identity [12]. Such biases may appear in several forms:

- **Generated Text Bias:** producing stereotypical, derogatory, or exclusionary content;
- **Ranking Bias:** privileging certain groups, narratives, or viewpoints;
- **Summarization Bias:** omitting perspectives or disproportionately framing events;
- **Translation Bias:** introducing gendered or cultural assumptions absent from the source text;
- **Recommendation Bias:** reinforcing popularity, conformity, or pre-existing inequalities.

These risks are amplified by scale [19]. Whereas earlier systems may have affected limited administrative decisions, LLMs can shape public discourse, educational access, professional workflows, and everyday reasoning across entire societies [19,21]. The size and opacity of such systems create conditions in which harmful biases may be difficult to predict, detect, or contest [10].

From Static Bias to Dynamic Bias

A major limitation in parts of the existing literature is the tendency to conceptualize bias primarily as an output problem: unequal predictions, unfair classifications, or objectionable generated content [7,8]. While these concerns remain essential, they may overlook deeper inferential structures through which systems acquire asymmetric knowledge about individuals and groups [7].

In many contemporary AI environments, harm arises not only from what a model explicitly states or decides, but from what it is capable of inferring [9,16]. Sensitive traits, vulnerabilities, preferences, health conditions, or social identities may be probabilistically reconstructed from weak signals, linguistic traces, or behavioral patterns long before any final output is produced [12,16]. Bias may therefore also reside in the differential capacity of systems to know more about some populations than others [9].

Toward a DBSD Perspective

The Deep Bias Systematic Deviation (DBSD) framework adds a complementary dimension to the literature on AI bias. Existing approaches often evaluate fairness through outputs, classifications, or representation disparities [7,8]. DBSD suggests that unfairness may emerge earlier—at the level of inferential capability itself [18]. If some groups are easier to profile, predict, or classify because their epistemic resistance is lower, then asymmetry exists prior to overt decision-making [9,16].

Under this view, bias is not only **unequal treatment**, but unequal inferability. Certain populations may become more transparent to algorithmic systems, while others remain comparatively opaque [9,16]. This shift from decision bias to knowledge-flow bias expands the ethical analysis of AI and offers a new basis for both measurement and regulation [7,18].

Concluding Observation

Contemporary scholarship convincingly shows that bias in AI is multi-causal, cumulative, and socially consequential. Yet as systems become increasingly inferential, adaptive, and deeply embedded in digital infrastructures, future research must move beyond visible outputs toward the hidden pathways through which knowledge is extracted, accumulated, and operationalized. It is precisely within this emerging space that the DBSD framework seeks to contribute.

The Deep Bias–Systematic Deviation (DBSD) Framework

The Deep Bias–Systematic Deviation (DBSD) framework proposes a new way of conceptualizing bias in contemporary artificial intelligence systems. Rather than treating bias as a binary property—biased versus unbiased—the framework defines bias as the degree of ease with which systematic deviation may be generated, amplified, transmitted, or inferred from existing data structures, model architectures, and decision environments [7].

Accordingly, the central analytical question shifts from:
"Is the system biased?"
to the more structurally relevant question:
"How easily can harmful deviation emerge, propagate, or intensify within the system?"
This shift is especially important in the age of machine learning and large language models, where explicit discriminatory rules are often absent, yet harmful disparities may still arise through latent correlations, proxy variables, optimization pressures, representational asymmetries, and inferential shortcuts [9,11,17,19]. In many modern systems, unfairness is produced less by overt design than by hidden computational pathways [16].

Conceptual Foundation: Bias as Flow Through a System

The DBSD framework draws an analogy from systems theory and electrical network logic. Just as current flows through circuits under resistance and impedance constraints, harmful bias may be understood as flowing through computational systems under structural conditions that either facilitate or inhibit its emergence.

Within this Analogy

- Bias Flow corresponds to the propagation of unequal or distorted outcomes;
- Structural Pressure corresponds to incentives, correlations, optimization objectives, or institutional demand;
- Resistance / Impedance corresponds to fairness constraints, transparency, auditing mechanisms, balanced data, and regulatory safeguards;
- Dissipated Harm corresponds to discriminatory or socially damaging outcomes.
- Under this view, some systems are highly conductive to bias, while others contain structural friction that inhibits harmful deviation.

This perspective allows bias to be analyzed dynamically rather than statically. Instead of asking only whether outputs are fair, DBSD asks how much resistance exists against the production of unfair outcomes.

Four-Layer Architecture of DBSD

The DBSD framework may be represented through four interconnected conceptual layers:

Input Layer

Observable signals entering the system, such as:

- Text, language prompts, and documents;
- Clicks and browsing behavior;
- Demographic variables;
- Geography and socioeconomic indicators;
- Interaction history;
- Institutional records.

These inputs may appear neutral while still containing latent proxies for sensitive traits.

Embedding Layer

Raw inputs are transformed into latent vector representations [21, 22]: $z=\phi(x)$ where x denotes the original input and $\phi(\cdot)$ is the learned representation function.

At this stage, explicit variables may disappear while hidden associations remain encoded geometrically [16].

Inference Network Layer

The system derives predictions, rankings, classifications, recommendations, or judgments through pathways that connect latent signals to downstream outcomes.

At this stage, stereotypes or group-level disparities may emerge indirectly through statistical dependence rather than direct instruction [9,16].

Impedance Layer

This layer captures the degree of difficulty required for biased outcomes to emerge. High impedance implies strong resistance to harmful deviation; low impedance implies structural susceptibility to bias generation.

Mathematical Formulation

Let: $p_i(x)$, denote the probability that input xproduces biased outcome i.

Overall bias exposure may be defined as: $B(x)=\sum_i w_i p_i(x)$ where: w_i are severity weights reflecting the normative importance of each harmful outcome. The impedance associated with bias channel iis defined as: $Z_i(x)=-\log(p_i(x))$

Thus:

- High probability of harmful bias →low impedance;
- Low probability of harmful bias →high impedance.

The aggregate DBSD score is: $DBSD(x)=\sum_i w_i \ Z_i(x)$

A higher DBSD score indicates greater systemic resistance to biased outcomes, whereas a lower DBSD score indicates greater structural vulnerability.

Interpretation

The primary innovation of DBSD lies in shifting analysis from static fairness metrics toward dynamic bias generation.

Traditional fairness literature often asks [8,17]:

- Are error rates equal?
- Are outcomes proportionate?
- Are groups treated similarly?
- These remain important questions. However, DBSD asks a deeper prior question:
- How easily can unequal outcomes emerge before they become visible?

This distinction is especially relevant in LLM-based systems, where harm may arise during internal inference long before final output is produced [11,19].

Forexample: "Prompt"→"Embedding"→"Attention"→"Output"

Bias may originate in hidden representational or inferential stages rather than solely in the final response.

Regulatory and Engineering Implications

The DBSD framework has practical implications for both technical design and public governance.

Engineering Implications

Developers may seek to increase bias impedance through [24,25]:

- Balanced and documented datasets;
- Debiasing embeddings;
- Fairness-aware objectives;
- Adversarial robustness methods;
- Transparent auditing systems;
- Monitoring of downstream disparities.

Regulatory Implications

Policymakers may use DBSD logic to evaluate not only discriminatory outputs, but also whether systems are structurally predisposed to generate harmful bias at scale [21,26].

This supports a transition from reactive fairness regulation toward preventive structural regulation.

Concluding Insight

The DBSD framework transforms bias from a purely descriptive label into a measurable systems property. It conceptualizes unfairness not only as what has already happened, but as what a system is structurally capable of producing.

In this sense, DBSD shifts fairness analysis from visible outcomes to latent capacities, from static metrics to dynamic pathways, and from ex post correction to ex ante prevention. It therefore provides a new conceptual foundation for measuring, mitigating, and governing bias in the era of intelligent systems.

Embedding Spaces, Inference, and the DBSD Framework

According to Friedler, fairness is not merely a mathematical property that can be inserted into an algorithm [18]. Rather, fairness depends on assumptions regarding the relationship between the **observed world,** the **true world**, and the **ideal world**. Because these worlds are frequently misaligned, many fairness objectives become difficult—and in some cases impossible—to satisfy simultaneously [27]. Different fairness

metrics therefore reflect different normative assumptions rather than universally valid technical truths [7,17].

The Deep Bias–Systematic Deviation (DBSD) framework extends this insight into the domain of embedding spaces and contemporary AI inference systems. In large language models and related architectures, embedding spaces function as latent representational environments in which words, concepts, entities, and behavioral signals are encoded as vectors in high-dimensional space [28,29]. Distances, directions, and angular relations among vectors often capture semantic proximity, statistical co-occurrence, contextual similarity, and functional association [28].

For example, the vector representation of doctor may appear close to hospital, patient, and medicine. Likewise, terms associated with residential location may become statistically connected to income, education level, ethnicity, or political orientation [16]. Such relationships may be useful for prediction and language understanding, yet they also create pathways through which bias may emerge indirectly [15,16].

The central concern is that embedding structures can encode sensitive correlations even when explicit labels are absent [15,16]. A model may infer protected attributes, social stereotypes, or group membership through geometric proximity rather than direct instruction [16,12]. Consequently, bias in modern AI systems often arises not only from visible variables, but from hidden semantic organization within latent spaces [16].

Within the DBSD framework, embedding spaces may therefore be understood as conductive layers through which inferential bias can travel. Information does not move randomly; it flows through structured semantic pathways that connect observable inputs to hidden attributes and ultimately to unequal outcomes [16].

This process may be represented schematically as:
"User Input"→"Embedding Space"→"Latent Relations"→"Hidden Attributes"→"Unequal Outcomes"

Large language models operate fundamentally as inference engines over such representational spaces [11,19]. Their outputs often reflect not only explicit prompts, but also latent relational structures learned from massive corpora [10,16].

The principal contribution of the DBSD framework is to shift analytical attention from visible discrimination alone toward the underlying mechanisms of inferential bias generation. Rather than asking only whether final outputs are unequal, DBSD asks how easily unequal outcomes can emerge from semantic geometry, relational clustering, and latent representational pathways [7,17].

Accordingly, embedding spaces should not be viewed as neutral technical containers [16]. They are socially consequential infrastructures in which meaning, probability, and power may become mathematically intertwined [12]. In this sense, DBSD provides a bridge between fairness theory, representation learning, and the governance of AI systems.

The Tri-World Embedding Bias Model: Mathematical Development and Numerical Illustration

Conceptual Motivation

Building on Friedler distinction between the observed world, the true world, and the ideal world, this section proposes a geometric extension for large language models [17]. The central idea is that each world can be represented as a vector, or as a centroid of vectors, within the embedding space of an LLM [28,29].
The model therefore distinguishes between three representational layers:

The Ideal World

A normative, bias-minimized world derived from liberal-democratic principles such as:

- Equality,
- Dignity,
- Non-discrimination,
- Inclusion,
- Fairness,
- Equal opportunity [29,30].

The True World

The actual social world in which empirically documented forms of bias exist, including:

- Gender discrimination,
- Racial discrimination,
- Skin-tone bias,
- Minority exclusion,
- Class inequality,
- Unequal institutional treatment [16,12].

The Observed World

The world is encoded in datasets and learned by the model. This includes representational patterns absorbed from training corpora, historical records, internet texts, user interactions, and institutional data [11].

The key claim is that bias may be modeled as a **geometric deviation** among these three worlds.

Vector Representation of the Three Worlds

Let: v_I denote the vector representation of the **Ideal World**, v_T denote the vector representation of the True World,

And v_O denote the vector representation of the **Observed World** encoded by the LLM.
Each vector may be constructed as the centroid of a set of embeddings [28].

For example: $v_T = \frac{1}{m}\sum_{j=1}^{m} b_j$ where b_j are embeddings representing empirically documented social biases [20, 24]. Finally: $v_O = \frac{1}{r}\sum_{i=1}^{r} o_i$, where o_i are embeddings extracted from model outputs, prompts, completions, training samples, or internal hidden states [11].

Cosine Similarity as a Measure of Bias Alignment

Cosine similarity measures angular proximity between two vectors: $\cos(u, v) = \frac{u \cdot v}{\|u\|\|v\|}$ Within the Tri-World model, three central similarities are of interest.

(a) Observed–Ideal Similarity $\cos(v_O, v_I)$ Measures how closely the model aligns with the normative ideal. A high value indicates proximity to fairness-oriented principles.
(b) Observed–True Similarity $\cos(v_O, v_T)$ Measures the extent to which the model reflects real-world bias structures.
(c) True–Ideal Similarity $\cos(v_T, v_I)$ Measures the moral distance

between existing social reality and the normative ideal. A low value indicates that society itself remains substantially removed from fairness standards.

Bias Deviation Measures

Observed Bias Deviation D_{OI}=1-cos(v_O,v_I) Measures how far the model is from the ideal world.

Social Bias Gap D_{TI}=1-cos(v_T,v_I) Measures how far the real social world is from the normative ideal [31, 32].

Model Reflection Bias ROT=cos(v_O,v_T) Measures the degree to which the model reproduces

The DBSD Bias Index

The Deep Bias–Systematic Deviation Index may be defined as:$DBSDI=\alpha D_{OI}+\beta R_{OT}$ where:

- α represents the normative importance of deviation from the ideal world [29];
- β represents the importance assigned to reproducing real-world bias [17];
- $\alpha+\beta=1$. And $\alpha,\beta\geq 0$

If: α=β=0.5

Then the model penalizes equally: distance from fairness ideals, and alignment with documented real-world bias. A high DBSDI score indicates that the model is both far from the ideal world and strongly aligned with biased empirical reality [12,16].

Numerical Example 1: Three-World Bias Measurement

Assume the following simplified 3-dimensional embedding vectors:v_I=(0.80,0.55,0.20),v_T=(0.30,0.85,0.43), v_O=(0.40,0.78,0.48)

These are illustrative vectors. The ideal vector emphasizes fairness and equality. The true-world vector reflects documented social bias. The observed vector reflects what the LLM has learned.

Step 1: Compute D_{OI} $v_O \cdot v_I$=(0.40)(0.80)+(0.78)(0.55)+(0.48)(0.20) =0.32+0.429+0.096=0.845

Norms: $\| v_O \| = \sqrt{0.40^2 + 0.78^2 + 0.48^2} = \sqrt{0.160 + 0.6084 + 0.2304} = \sqrt{0.9988} = 0.9994$

$$\| v_I \| = \sqrt{0.80^2 + 0.55^2 + 0.20^2} = \sqrt{0.64 + 0.3025 + 0.04} = \sqrt{0.9825} = 0.9912$$

Thus: $\cos(v_O, v_I) = \frac{0.845}{(0.9994)(0.9912)} = 0.853$

Therefore: D_{OI}=1-0.853=0.147

The observed model representation has a bias deviation of 0.147 from the ideal world.

Step 2: Compute R_{OT} $v_O \cdot v_T$=(0.40)(0.30)+(0.78)(0.85)+(0.48)(0.43) =0.120+0.663+0.2064=0.9894

Norm of

v_T: $\| v_T \| = \sqrt{0.30^2 + 0.85^2 + 0.43^2} = \sqrt{0.09 + 0.7225 + 0.1849} = \sqrt{0.9974} = 0.9987$

Thus: R_{OT}=cos(v_O,v_T)

This means that the observed model representation is highly aligned with the biased true-world representation.

$$= 0.991$$

Step 3: Compute D_{TI} $v_T \cdot v_I$=(0.30)(0.80)+(0.85)(0.55)+(0.43)(0.20)=0.240+0.4675+0.086=0.7935

$$\cos(v_T, v_I) = \frac{0.7935}{(0.9987)(0.9912)} = 0.802$$

Therefore:D_{TI}=1-0.802=0.198

The true social world itself is significantly distant from the ideal world.

Step 4: Compute DBSDI

Let:α=0.5,β=0.5Then:$DBSDI$=0.5(1-D_{OI})+0.5(R_{OT})=0.5(0.147) +0.5(0.991)=0.0735+0.4955=0.569This indicates a relatively elevated bias condition: the model is insufficiently close to the ideal world while remaining strongly aligned with empirical bias structures. Interpretation: The model has a relatively high DBSDI because it is close to the biased true world and insufficiently close to the ideal world.

Bias Reduction Through Vector Steering

Bias mitigation may be modeled as controlled movement of the observed vector toward the ideal vector [19]:$v'_O=v_O+\lambda(v_I-v_O)$, where:$0\leq\lambda\leq 1$ and:

- λ=0: no correction;
- λ=1: full movement to the ideal vector;
- $0<\lambda<1$: partial correction.

This represents a geometric intervention in embedding space [15,16].

Numerical Example 2: Bias Reduction

Using:v_0=(0.40,0.78,0.48), v_I=(0.80,0.55,0.20)

Let: λ=0.40. Then: $vO'=vO+0.40(vI-vO)$ First compute:v_I-v_0 =(0.80−0.40,0.55−0.78,0.20−0.48) =(0.40,−0.23,−0.28) Then: $0.40(v_I-v_0)$=(0.16,−0.092,−0.112) Therefore: v_0' =(0.40,0.78,0 .48)+(0.16,−0.092,−0.112) =(0.56,0.688,0.368)

Step 1: New Observed–Ideal Similarity

$v'_0 \cdot v_I$=(0.56)(0.80)+(0.688)(0.55)+(0.368)(0.20) =0.448+0.3784+0.0736=0.900

Norm:$\| v'_O \| = \sqrt{0.56^2 + 0.688^2 + 0.368^2} = \sqrt{0.3136 + 0.4733 + 0.1354}$

$$= \sqrt{0.9223} = 0.9604$$

Thus:$\cos(v_0', v_I) = \frac{0.900}{(0.9604)(0.9912)} = 0.945$, $D'_{OI} = 1 - 0.945 = 0.055$

Bias deviation was reduced from: 0.147→0.055 This is a reduction of: $\frac{0.147-0.055}{0.147} = 0.626$ or approximately 62.6%.

Step 2: New Observed–True Similarity

$v_O' \cdot v_T$=(0.56)(0.30)+(0.688)(0.85)+(0.368)(0.43) =0.168+0.5848+0.1582=0.911

Thus: $R'_{OT} = \frac{0.911}{(0.9604)(0.9987)} = 0.950$

The alignment with the biased true world decreased: 0.991→0.950

This is a reduction of: $\frac{0.991-0.950}{0.991}$=0.0413 or approximately **4.13%.**

This indicates partial reduction in the model's reproduction of real-world bias.

Step 3: New DBSDI $DBSDI'$=0.5(1-D'OI)+0.5(R'OT) =0.5(0.055)+0.5(0.950) =0.0275+0.475=0.5025

The index decreased: 0.569→0.503 This is a reduction of: $\frac{0569-0.503}{0.569} = 0.116$ or approximately 11.6%. This represents measurable bias reduction.

Stronger Correction Example
Let:λ=0.70

Then:$v_0''=v_0+0.70(v_I-v_0)=(0.40,0.78,0.48)+0.70(0.40,-0.23,-0.28)=(0.40,0.78,0.48)+(0.28,-0.161,-0.196)=(0.68,0.619,0.284)$Compute:$v_0''\cdot v_I=(0.68)(0.80)+(0.619)(0.55)+(0.284)(0.20)=0.544+0.3405+0.0568=0.9413$
Norm:

$\| v_O'' \| = \sqrt{0.68^2 + 0.619^2 + 0.284^2} = \sqrt{0.4624 + 0.3832 + 0.0807} = \sqrt{0.9263} =$

0.9624 Thus: $\cos(v_O'', v_I) = \frac{0.9413}{(0.9624)(0.9912)} = 0.986$

Bias deviation: $D''_{0I}=1-0.986=0.014$

This is a strong reduction from the original: 0.147→0.014 or approximately 90.5% reduction.

Proposition
Let $v_O$be the observed vector and v_I the ideal vector. Define:$v_O'=v_O+\lambda(v_I-v_O)$
with $0<\lambda\leq1$. If movement occurs along the segment from v_O toward v_I, then angular distance decreases and: $\cos(v_O',v_I)\geq\cos(v_O,v_I)$, Therefore:$D_{OI}'\leq D_{OI}$ That is, movement toward the ideal vector reduces bias deviation. **Interpretation**
A DBSD correction increases alignment with the ideal world and reduces bias deviation.

Interpretation for LLM Bias Mitigation
The mathematical framework suggests that LLM bias mitigation can be modeled as controlled steering in embedding space [19, 20]. A model is biased when its observed representation:

1. Is distant from the ideal vector; and
2. Is overly aligned with empirically biased social reality.

Bias Mitigation therefore requires:
- Increasing similarity to the ideal world: $\cos(v_O',v_I)>\cos(v_O,v_I)$
- Decreasing alignment with biased social reality: $\cos(v_O',v_T)<\cos(v_O,v_T)$
- Reducing the DBSDI score. Thus *DBSDI'<DBSDI*

This transforms fairness intervention into a measurable geometric optimization problem.

Concluding Insight
The Tri-World Embedding Bias Model converts bias analysis from a purely qualitative or output-based discussion into a measurable representational framework. It allows researchers to ask not only whether a model produces biased outputs, but where its internal representations are positioned relative to [7,17]:

1. The normative ideal world,
2. The empirically biased true world, and
3. The observed world encoded by data.

In this sense, the model provides a mathematical bridge between social theory, moral philosophy, embedding geometry, and AI governance [29,30].

Regulatory Functions in the Three-World /DBSD Framework
Conceptual Role of Regulation
Within the DBSD (Deep Bias–Systematic Deviation) framework, regulation performs three central functions.

(1) Reducing Bias in the Observed World Regulation seeks to reduce distortions in the observed world—namely, the world represented by AI systems, datasets, and large language models [21,26]. This may be achieved through legal standards, institutional safeguards, transparency obligations, fairness auditing, documentation requirements, and anti-discrimination norms [21,24,25].

2) Determining the Normative Weight of Ideal Alignment
Regulation determines the importance assigned to the gap between the observed world and the ideal world through the parameter α. The parameter α measures how strongly society values alignment with fairness, equality, dignity, inclusion, and non-discrimination [28,29]. A higher value of α implies stricter normative expectations.

(3) Determining the Weight of Realism
Regulation also determines the importance assigned to the relationship between the true world and the observed world through the parameter: β. Where β measures how strongly society values descriptive realism, empirical accuracy, and representation of existing social conditions [17]. A higher value of β implies greater tolerance for reflecting real-world inequalities in the name of realism.

Why Regulation Matters
Biases in the observed world may become informational foundations for highly consequential decisions, including [9,26]:
- Law and judicial processes;
- Hiring and career opportunities;
- Access to education;
- Credit, insurance, and economic rights;
- Healthcare allocation;
- Political communication;
- Distribution of public resources.

Because AI-mediated representations increasingly shape institutional judgment, reducing the gap between the observed world and the ideal world is not merely a technical preference, it is a democratic and ethical necessity [21,26].

Regulatory Control Through Corrective Intervention
Within the DBSD framework, regulatory intervention may be represented by the parameter:λ where λ denotes the strength of corrective pressure applied to move the observed-world representation toward the ideal-world representation.
Thus:
- λ=0: no intervention;
- $0<\lambda<1$: partial intervention;
- λ=1: maximal intervention.

The corrected observed vector is defined as: $v_O'=v_O+\lambda(v_I-v_O)$ where:
- v_O= observed-world vector,
- v_I= ideal-world vector.

This formalizes regulation as geometric steering of representational systems rather than solely punishment after harm occurs [21,26].

Mathematical Regulatory Model
Corrected Observed Vector Let: v_O= observed world vector, v_I= ideal world vector
Regulatory correction is:$v_O'=v_O+\lambda(v_I-v_O)$, where:$0\leq\lambda\leq1$, $\lambda=0$: no intervention, $\lambda=1$: full shift to ideal world

DBSD Objective Function We define total bias cost as: $J(\lambda)=\alpha(1-\cos(v_O',v_I))+\beta\cos(v_O',v_T)$ where: first term = distance from ideal world, second term = similarity to biased true world [15,31].

Regulatory Condition

If: $\alpha = 2\beta$. Thus:$\beta = \frac{\alpha}{2}$, Substitute: $J(\lambda) = \alpha(1-\cos(v_O', v_I)) + \frac{\alpha}{2}\cos(v_O', v_T)$

Factor out α: $J(\lambda) = \alpha\left[(1-\cos(v_O', v_I)) + \frac{1}{2}\cos(v_O', v_T)\right]$

So minimizing J(λ) is equivalent to minimizing:

$$F(\lambda) = (1-\cos(v_O', v_I)) + \frac{1}{2}\cos(v_O', v_T)$$

Required Formula for λ The regulator should choose

$$\lambda^* = \arg\min_{0\leq\lambda\leq1}\left[1-\cos\left(v_O+\lambda(v_I-v_O), v_I\right) + \frac{1}{2}\cos\left(v_O+\lambda(v_I-v_O), v_T\right)\right]$$

Approximate Closed-Form Practical Formula
If vectors are normalized and geometry is smooth, a practical approximation is the **Final Formula:**

$$\lambda^* = \frac{2(1-\cos(v_O,v_I))}{2(1-\cos(v_O,v_I))+\cos(v_O,v_T)}$$ under the condition: $\alpha=2\beta$

Interpretation
If the observed world is:
Far from the ideal world: $\cos(v_O,v_I)\downarrow$,then: $\lambda^*\uparrow$, means stronger intervention.
Too close to the biased true world: $\cos(v_O,v_T)\uparrow$, then: $\lambda^*\uparrow$, means stronger intervention.

Interpretation of Parameters
If society prioritizes justice and equality [31], then:$\alpha>\beta$
If society prioritizes descriptive realism [15], then:$\beta>\alpha$
If both values are equal:$\alpha=\beta$. The model adopts a balanced position between normative correction and empirical reflection.

This feature makes the DBSD framework adaptable across jurisdictions, political cultures, and regulatory philosophies [30].

Regulator Trash Hold
Suppose regulators define a maximum acceptable level of observed bias:γ where:$0\leq\gamma\leq1$. Observed bias is defined as:D_{OI}= $1-\cos(v_O,v_I)$
Regulatory intervention is triggered whenever:$D_{OI}>\gamma$
Thus:
- low γ: strict regulation;
- medium γ: moderate regulation;
- high γ: permissive regulation.

This allows policymakers to formalize different tolerance levels for AI bias [21,26].

Here, λ represents the regulatory correction strength. The regulator should select the admissible solution: $0\leq\lambda^*\leq1$ Thus, λ*represents the minimal intervention strength required to ensure that the observed-world bias does not exceed the permitted threshold γ. Let:$c=\cos(v_O,v_I)$. Assuming that v_O and v_I are normalized vectors, the minimum value of λ required to satisfy the regulatory threshold is obtained by

solving: $\cos(v_O+\lambda(v_I-v_O),v_I)=1-\gamma$ .This yields the quadratic solution: $\lambda^* = \frac{-B+\sqrt{B^2-4AC}}{2A}$

where:$A=(1-c)((1-c)-(1-\gamma)^2)$, $B=2(1-c)(c-(1-\gamma)^2)$,$C=c^2-(1-\gamma)^2$
The regulator should select the admissible solution: $0\leq\lambda^{\wedge}*\leq1$ Thus, λ^*represents the minimal intervention strength required to ensure that the observed-world bias does not exceed the permitted threshold γ.

Special Case: γ=0.5
If the permitted bias threshold is:$\gamma=0.5$ then:$1-\gamma=0.5$ and the regulatory condition becomes:$\cos(v_O',v_I)\geq0.5$ In this case, the above formula reduces to the threshold condition previously discussed.

Numerical Examples
Suppose: $\cos(v_O,v_I)=0.72$, $\cos(v_O,v_T)=0.88$ Then: $\lambda^* = \frac{2(1-0.72)}{2(1-0.72)+0.88} = \frac{0.56}{1.44} = 0.389$ So regulator should impose:$\lambda=0.39$ meaning a 39% corrective shift toward the ideal world.

Stronger Example

If: $\cos(v_O, v_I) = 0.45$, $\cos(v_O, v_T) = 0.91$ Then: $\lambda^* \frac{2(1-0.45)}{2(1-0.45)+0.91} = \frac{1.10}{2.01} = 0.547$

Need **54.7% correction**.

Interpretation When the stronger the deviation of AI-generated observed reality from the ideal normative world, and the stronger its alignment with empirical discriminatory structures, the stronger the regulatory correction parameter λ should be. That means regulators may choose different acceptable bias thresholds. Because $\cos(v_O',v_I)\geq1-\gamma$ we get:

Table 1: Threshold Meaning

$\gamma=0.2$	very strict regulation
$\gamma=0.5$	moderate regulation
$\gamma=0.7$	permissive regulation

Dynamic Regulation in the Three-World / DBSD Framework
We now extend the model by assuming that the regulator does not determine the correction parameter λonly once but continuously updates it over time in response to measured bias levels [26]. Thus, regulatory intervention becomes a time-dependent function: λ(t) where tdenotes time. Such adaptive regulatory mechanisms align with emerging governance frameworks that emphasize continuous monitoring, risk-based intervention, and lifecycle oversight in AI systems [21,26].

Time-Dependent Bias Measurement
At each time step t, the bias between the observed world and the ideal world is defined as: $D_{OI}(t)=1-\cos(v_O(t),v_I(t))$, where: $v_O(t)$

is the observed-world representation at time t, vI (t)is the ideal-world representation at time *t*.
This allows both the observed world and the normative ideal to evolve dynamically.
Let $0 \le \gamma \le 1$ be a general regulatory threshold

Regulatory Threshold Rule
Assume that the regulator sets a maximum permitted bias level: $B_{max}=\gamma$ The intervention rule is then:

$$\lambda(t) = \begin{cases} 0, & D_{OI}(t) \le \gamma \\ \lambda^*(t), & D_{OI}(t) > \gamma \end{cases}$$

This means:
- If the measured bias remains below the threshold, no intervention is required.
- If the measured bias exceeds the threshold, corrective intervention is activated.

Continuous Dynamic Update Rule
A smoother regulatory mechanism can be modeled by continuously adjusting λ rather than switching discretely: $\lambda(t+1)=\lambda(t)+\eta(D_{OI}(t)-D_{max})$
where: η is the regulator's response rate, $D_{OI}(t)$ is the observed bias at time *t*, $BD_{max}=\gamma$ is the tolerated bias threshold. Interpretation If: $D_{OI}(t)>\gamma$ then: $\lambda(t+1)>\lambda(t)$, meaning regulatory pressure increases.
If: $D_{OI}(t)<\gamma$ then: $\lambda(t+1)<\lambda(t)$, meaning intervention may be relaxed.
Thus, regulation behaves as an adaptive feedback controller.

Bounded Regulation Constraint
To ensure that the correction parameter remains within valid limits: $0 \le \lambda(t) \le 1$
we define: $\lambda(t+1)=\min\{1,\max\{0,\lambda(t)+\eta(D_{OI}(t)-\gamma)\}\}$
This guarantees that: λ=0: no intervention, λ=1: maximum intervention

Long-Term Feedback Effect
Regulation may be modeled as a dynamic feedback mechanism [29] in which the correction parameter λ is adjusted according to the gap between the measured observed–ideal bias and the permitted threshold γ. When $D_{OI}(t)$ exceeds γ, the regulator increases intervention strength; when *DOI* (*t*)remains below γ, regulatory pressure may be reduced.

Numerical Example: Dynamic Regulation with γ=0.5
Consider the dynamic regulatory rule: $\lambda(t+1)=\min\{1,\max\{0,\lambda(t)+\eta(D_{OI}(t)-\gamma)\}\}$
where:
- γ=0.5 is the maximum permitted bias threshold,
- η=0.4 is the regulator's response rate,
- λ(0)=0is the initial intervention level.

Numerical Table 2

Time *t*	$B_{OI}(t)$	$B_{OI}(t)-\gamma$	Update Rule	$\lambda(t+1)$
0	0.70	0.20	0+0.4(0.20)	0.08
1	0.65	0.15	0.08+0.4(0.15)	0.14
2	0.55	0.05	0.14+0.4(0.05)	0.16
3	0.48	-0.02	0.16+0.4(-0.02)	0.152
4	0.42	-0.08	0.152+0.4(-0.08)	0.120
5	0.50	0.00	0.120+0.4(0)	0.120

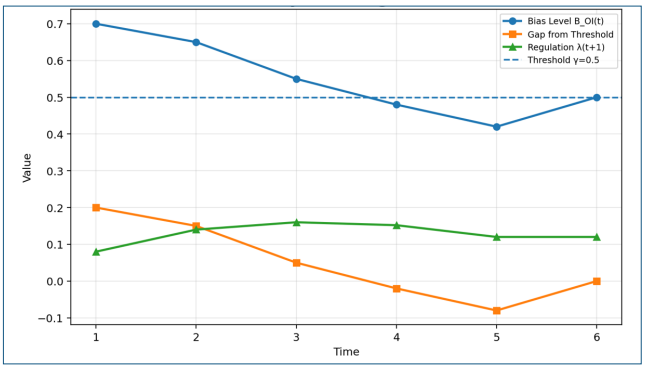

Figure 1: Dynamic Regulation Trends

Figure 1 presents the cleaned dynamic regulation graph. The figure illustrates how regulatory pressure increases when bias exceeds the threshold and stabilizes as the system approaches compliance.

In this example, the regulator sets the acceptable bias threshold at γ=0.5. Whenever the observed–ideal deviation exceeds this level, the correction parameter λincreases. Once the deviation falls below the threshold, intervention pressure is gradually reduced. This demonstrates how dynamic regulation may function as a self-correcting control process over AI-generated bias.

Feedback Effect on the True World
The broader purpose of regulation is not limited to correcting AI outputs in the observed world. By reshaping incentives, information flows, institutional practices, and decision norms, fairer observed representations may gradually influence the true world itself.

Over time, repeated interventions may reduce real social inequalities, thereby narrowing the gap between: v_T and v_I In this sense, regulation does not merely react to society—it may also help
The DBSD regulatory framework moves beyond traditional ex post enforcement. Rather than waiting for discriminatory outcomes to occur, it enables **preventive governance** by identifying when representational systems are drifting toward harmful deviation and by correcting them in advance. Thus, regulation becomes measurable, dynamic, and structurally integrated into AI governance transform society.

Conclusions
This article has introduced the Deep Bias–Systematic Deviation (DBSD) framework as a novel theoretical, mathematical, and regulatory approach to understanding bias in contemporary artificial intelligence systems. The central contribution of the framework lies in shifting the analytical focus from the question of whether a system is biased to the more structurally meaningful question of how easily bias can be generated, inferred, amplified, and translated into socially consequential outcomes. In an era increasingly shaped by large language models, latent representation systems, and automated decision infrastructures, such a shift is both conceptually necessary and practically urgent.

The DBSD framework advances several original contributions. First, it reconceptualizes bias as a **dynamic structural process** rather than a static output defect. While traditional approaches evaluate fairness through observable disparities, unequal treatment, or statistical imbalances, DBSD emphasizes the latent inferential pathways through which harmful deviation may arise prior to visible outcomes. This perspective reframes bias as a function of systemic susceptibility to inference, rather than solely as an observed property of decisions.

Second, the framework extends fairness analysis into **embedding spaces** through the Tri-World model, enabling systematic comparison among:

- the observed world represented by AI systems,
- the true world of empirical social reality, and
- the ideal world grounded in normative principles such as equality, dignity, and non-discrimination.

By representing these worlds geometrically, the model provides a unified structure for analyzing bias across philosophical, social, and computational domains.

Third, the article develops a set of **quantitative tools** based on cosine similarity, bias impedance, and controlled vector steering. These tools enable the measurement and mitigation of bias within high-dimensional representation spaces. Rather than relying solely on qualitative judgments or post hoc evaluation, DBSD introduces a formal and operationalizable framework for estimating and reducing representational deviation.

Fourth, the framework advances a **dynamic regulatory paradigm** in which intervention strength adapts over time through threshold-sensitive feedback mechanisms. By modeling regulation as a continuous control process governed by parameters such as λ and γ, DBSD transforms governance from a reactive system of ex post enforcement into a proactive system of adaptive intervention. In this sense, regulation becomes measurable, responsive, and structurally embedded within AI systems themselves.

More broadly, the DBSD framework suggests that effective AI governance cannot rely exclusively on auditing outputs after decisions are made. Instead, fairness must be addressed at earlier stages, including data construction, latent representation, inferential architecture, optimization design, and deployment environments. This shift from **output fairness to inference fairness** represents a fundamental reorientation of both technical and regulatory thinking.

The originality of the present study lies in its integration of multiple intellectual domains—including ethics, social theory, systems theory, vector geometry, machine learning, and regulatory design—into a single coherent analytical framework. By doing so, the article provides not only a diagnostic language for understanding bias, but also a constructive methodology for measuring, mitigating, and governing it. The DBSD paradigm opens several promising directions for future research, including:

- Explainable fairness in AI systems,
- Adaptive regulation of large language models,
- Governance of embedding representations,
- Bias auditing in high-dimensional inference systems,
- Comparative normative models across jurisdictions, and
- Democratic oversight of intelligent infrastructures.

Ultimately, the DBSD framework proposes that bias should be understood not merely as what a system has done, but as what a system is structurally capable of doing. This shift—from observed harm to latent capacity—may prove essential for the development of robust, accountable, and ethically aligned AI systems in the coming decades.

References

1. Oxford Learner's Dictionaries. Bias definition.
2. Merriam-Webster Dictionary. Bias definition.
3. Casella, G., Berger, R. (2024). Statistical Inference. CRC Press.
4. Tversky, A., Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases.
5. Solove, D. J. (2010). Understanding Privacy. Harvard UP.
6. Nissenbaum, H. (2009). Privacy in Context. Stanford UP.
7. Barocas, S., Hardt, M., Narayanan, A. (2023). Fairness and Machine Learning. MIT Press.
8. Hardt, M., Price, E., Srebro, N. (2016). Equality of Opportunity in Supervised Learning.
9. Barocas, S., Selbst, A. D. (2016). Big Data's disparate impact. California Law Review.
10. Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S. (2021). On the dangers of stochastic parrots. FAccT.
11. Kamath, U. (2024). Large Language Models: A Deep Dive. Springer.
12. Gallegos, I. (2024). Bias and fairness in LLMs: A survey. Computational Linguistics.
13. Oppenheim, Y. (2026a). Beyond Disclosure.
14. Oppenheim, Y. (2026b). Privacy as Epistemic Impedance.
15. Bolukbasi, T. (2016). Debiasing Word Embeddings. NeurIPS.
16. Caliskan, A., Bryson, J., Narayanan, A. (2017). Human-like biases in corpora. Science.
17. Kleinberg, J., Mullainathan, S., Raghavan, M. (2016). Inherent Trade-Offs in Fair Risk
18. Friedler, S. A. (2021). The (im)possibility of fairness.
19. Bommasani, R. (2021). Opportunities and Risks of Foundation Models. Stanford.
20. Weidinger, L. (2022). Taxonomy of Risks Posed by Language Models. FAccT.
21. European Union. (2024). AI Act.
22. OECD. (2019). OECD Principles on AI.
23. Sambasivan, N. (2021). Data Cascades in High-Stakes AI. CHI.
24. Gebru, T. (2021). Datasheets for Datasets. CACM.
25. Mitchell, M. (2019). Model Cards for Model Reporting. FAccT.
26. NIST. (2023). AI Risk Management Framework.
27. Devlin, J. (2019). BERT. NAACL.
28. Mikolov, T. (2013). Efficient Estimation of Word Representations.
29. OpenAI. (2023). GPT-4 Technical Report.
30. Ji, Z. (2023). Survey of Hallucination in NLG.