From Ethical Declarations to Provable Independence: An Ontology-Driven Optimal-Transport Framework for Certifiably Fair AI Systems

Sukriti Bhattacharya[1] and Chitro Majumdar[2]

[1] Senior Scientist, Trustworthy AI, Luxembourg Institute of Science & Technology, Maison de l'innovation 5, L-4362, Luxembourg
sukriti.bhattacharya@list.lu
[2] Chief Investment Risk Strategist for Sovereign Institutions
Founder, RsRL, Jumeirah Beach Residence (JBR), P.O.Box 29215, Dubai, United Arab Emirates
chitro.majumdar@rsquarerisklab.com

Abstract. This paper introduces a novel framework for achieving provably fair AI systems that addresses the fundamental limitation of existing bias mitigation approaches: their inability to systematically identify and eliminate all forms of sensitive information, including subtle proxies that enable discriminatory outcomes through seemingly neutral features. Our integrated methodology leverages ontology engineering with OWL 2-QL to formally declare sensitive attributes and systematically discover their proxies through logical inference, compiling these into a comprehensive σ-algebra G that represents the complete structure of biased measurable patterns. We then apply Delbaen–Majumdar optimal transport theory to construct new variables that are provably independent of G while minimizing L^2-distance to preserve predictive accuracy. This approach transcends correlation-based debiasing techniques by providing exact independence guarantees rather than mere decorrelation. The framework's theoretical foundation rests on three key innovations: formalizing algorithmic bias as a structural property of learned σ-algebras, demonstrating systematic compilation of ontological knowledge into measure-theoretic structures, and establishing optimal transport as the unique solution for fair representation learning. Through detailed analysis of loan approval systems, we illustrate how ontology-guided σ-algebra construction captures indirect bias pathways (e.g., ZIP codes as proxies for race) that traditional methods overlook, ensuring comprehensive fairness while maintaining decision utility. The resulting methodology delivers a unified, certifiable approach to algorithmic fairness with applications spanning bias-free lending decisions and context-aware language model outputs, advancing the field from heuristic bias mitigation toward mathematically grounded fairness guarantees essential for trustworthy AI deployment in high-stakes societal applications.

Keywords: AI Fairness · Algorithmic Bias · Ontology Engineering · Sigma-Algebra

Acknowledgment

The authors gratefully acknowledge Professor Emeritus ***Dr. Freddy Delbaen*** *from the Department für Mathematik, ETH Zürich, Rämistrasse 101, 8092 Zürich, for his invaluable help, unwavering support, and insightful discussions and comments that significantly contributed to the development of this work.*

1 Introduction

Artificial intelligence (AI) systems, encompassing machine learning models and large language models (LLMs), are increasingly integral to high-stakes decision-making in domains such as financial lending, hiring, criminal justice, healthcare, and content generation. Yet, these systems frequently perpetuate algorithmic bias, producing unfair outcomes by learning and reproducing undesirable patterns embedded in historical training data. Such biases often stem from datasets that reflect societal inequities, where features like socio-economic status, race, gender, or their proxies (e.g., zip codes or employment history) inadvertently shape model decisions. For instance, automated loan approval systems trained on historical lending data may disproportionately deny applications from lower-income or minority groups, or offer them less favorable terms, despite comparable financial profiles. Similarly, as demonstrated by [16], LLMs like Google Gemini have generated historically inaccurate depictions, such as racially diverse Nazis, while models like Claude have erroneously asserted uniform military fitness standards across genders. These errors highlight a critical flaw termed "difference unawareness," where fair differentiation is conflated with harmful prejudice, leading to contextually inappropriate outcomes.

This issue is not merely technical but carries profound ethical and societal implications, demanding urgent and rigorous intervention. Algorithmic bias exacerbates existing disparities, undermines public trust in AI systems, and engenders severe ethical and legal consequences, including the perpetuation of systemic inequities and violations of human dignity. In practice, biased AI outcomes can deny individuals equitable access to critical opportunities—such as loans, jobs, or medical care—while entrenching modern forms of discrimination through ostensibly neutral "color-blind" or "gender-blind" approaches [16]. For example, enforcing uniform treatment in AI systems can backfire by ignoring contextual necessities, such as legal distinctions (e.g., gender-specific military drafts) or differential harm assessments (e.g., derogatory terms carrying greater weight against marginalized groups). These failures not only worsen societal outcomes but also erode the promise of AI as a force for progress.

Addressing algorithmic bias is therefore paramount to ensuring fairness, promoting inclusivity, and aligning AI systems with ethical standards and regulatory frameworks. By developing principled methodologies that eliminate dependence on sensitive attributes while embracing *contextual difference awareness*, we can create AI systems that deliver equitable, contextually sensitive, and trustworthy outcomes across diverse societal landscapes. Such an approach prevents AI from

becoming a vector for division, instead harnessing its potential to advance societal good. To tackle this challenge, we propose a novel framework that integrates ontology engineering with measure-theoretic sigma-algebras to achieve provably fair AI systems. This approach conceptualizes bias as a structural property of the model's learned σ-algebra–the mathematical structure that encodes the "measurable" patterns and correlations derived from training data. By leveraging the declarative power of ontologies, we formally define sensitive attributes and their proxies (e.g., race, gender, or zip code) within a logical framework, specifically using OWL 2-QL for its computational efficiency and scalability in query answering [9]. These ontological axioms are compiled into a sigma-algebra G, which represents the undesirable information the model should not rely upon. Subsequently, we employ optimal transport techniques, building on advancements by [4,5,3,15], to construct a new random variable Y that is independent of G and minimally distant from the original biased variable in the L^2-norm. This ensures that the resulting representations are both fair (by achieving exact independence from sensitive attributes) and accurate (by preserving maximal information from the original data).

The remainder of this paper is structured as follows. Section §2 reviews the state of the art on algorithmic bias mitigation, ontology-based frameworks, and measure-theoretic approaches, situating our contribution within ongoing debates on fairness. Section §3 introduces the σ-algebra formalism, establishing bias as a sub-σ-algebra and defining fairness through independence. Section §4 develops the ontological foundation, presenting how sensitive attributes and their proxies are formally encoded. Section §5 integrates these strands into the notion of ontology-guided σ-algebras, demonstrating their robustness and applying them to the loan approval case study. Section§6 outlines an implementation blueprint, mapping the theoretical framework into a reproducible pipeline for practice and auditability. Finally, Section §7 concludes with key insights, emphasizing the theoretical and practical implications of our approach and identifying avenues for future research.

2 Literature Review

The growing deployment of artificial intelligence (AI) systems in high-stakes domains—such as financial lending, hiring, criminal justice, healthcare, and content generation—has amplified concerns about algorithmic bias, where models learn and perpetuate undesirable patterns from biased training data [2]. These biases often stem from historical inequalities reflected in datasets, manifesting through sensitive attributes like race, gender, socio-economic status, or proxies (e.g., zip codes or employment history), leading to unfair outcomes that exacerbate societal disparities [10,14]. For instance, automated loan approval systems may disproportionately disadvantage marginalized groups, while clinical prediction models can exhibit fairness drift over time, widening performance gaps across subpopulations like race and sex [5,4,2]. Traditional bias mitigation strategies have primarily focused on correlation-based techniques, such as

dataset balancing, fairness constraints, adversarial learning, and reweighting [10]. However, these approaches often require group annotations, which are costly or unavailable, and may inadvertently reduce overall model accuracy or fail to address deeper structural biases [10,18]. Recent critiques highlight the limitations of "difference-unaware" fairness, which enforces uniform treatment but overlooks contextual distinctions, potentially worsening inequities in applications requiring group-specific awareness [16]. Moreover, emerging issues like fairness drift—where post-deployment model fairness degrades due to evolving data distributions or clinical practices—underscore the need for ongoing monitoring and adaptive methods [2]. In clinical settings, for example, random forest models for post-surgical outcomes have shown increasing gaps in accuracy and specificity between demographic groups over a decade, with model updates sometimes exacerbating rather than alleviating these disparities [2].

Ontology-based frameworks have gained traction for standardizing AI concepts and addressing ethical concerns, including bias. The Artificial Intelligence Ontology (AIO) provides a comprehensive hierarchy of AI methodologies, with a dedicated Bias branch encompassing 61 subclasses drawn from sources like NIST reports, categorizing biases across the AI lifecycle (e.g., computational, historical, societal) to promote fairness and transparency [11]. Similarly, knowledge graph-based ontological frameworks integrate predictive analytics, AI, and big data for decision-making systems, emphasizing ethical considerations such as algorithmic bias mitigation through diverse datasets, explainable AI (XAI), and compliance with regulations like the EU AI Act [3]. These ontologies facilitate modular composition and ethical evaluations, aiding in bias identification and responsible AI deployment in sectors like healthcare and public policy [11,14]. Data-centric debiasing methods represent another evolving strand, focusing on identifying and removing problematic samples or shortcuts without extensive annotations. Data Debiasing with Datamodels (D3M) uses datamodeling to pinpoint training examples harming subgroup robustness, improving worst-group accuracy across datasets like CelebA and Waterbirds while preserving natural accuracy [10]. Complementing this, Shortcut Hull Learning (SHL) formalizes shortcuts in a probabilistic measure-theoretic space, defining a "shortcut hull" as the minimal set of unintended features and employing diverse model suites to diagnose and eliminate them for unbiased evaluations [18]. SHL's framework, validated on topological datasets, reveals insights into model capabilities (e.g., CNNs outperforming Transformers in global reasoning) and ensures independence from confounding correlations, enhancing fairness in high-dimensional AI assessments [18]. Building on these foundations, measure-theoretic approaches offer a principled path to achieving statistical independence from biased structures. Our prior works have pioneered this direction: [4] and [5] leverage optimal transport and Monge-Kantorovich problems to construct random variables independent of a sub-σ-algebra representing biased information, minimizing L^2-distance while ensuring stronger guarantees than mere decorrelation. Extending this, [3] develops practical algorithms for discrete spaces, with applications to eliminating socio-economic dependencies in mortgage underwriting. In parallel, [18] concep-

tualizes bias as a property of learned σ-algebras, proposing independence-based constructions to unify fair representation learning, counterfactual reasoning, and differential privacy.

Despite these advancements, significant gaps persist. Ontology frameworks, while useful for standardization, often lack integration with dynamic monitoring for fairness drift [11,2]. Data debiasing methods like D3M and SHL excel in annotation-scarce settings but may not scale to multidimensional cases or guarantee long-term independence from evolving biases [10,18]. Measure-theoretic methods provide theoretical rigor but require further empirical validation in real-world AI pipelines [4]. This review highlights the need for hybrid approaches that combine ontological structuring, data-centric debiasing, drift monitoring, and measure-theoretic independence to foster equitable, sustainable AI systems.

3 The σ-Algebra: A Formalization of Observable Information

Probability Space and Measurable Events A *probability space* is a triple $(\Omega, \mathcal{F}, \mathbb{P})$, where:

- The **sample space** Ω is a non-empty set of all possible outcomes. In our context, this can be a dataset of individual data points.
- $\mathcal{F}$ is a **σ-algebra** on Ω, which represents the collection of all events about which we can reason probabilistically.
- $\mathbb{P}$ is a **probability measure** on $\mathcal{F}$, a function that assigns a likelihood (a value between 0 and 1) to each event in $\mathcal{F}$.

3.1 The σ-Algebra: A Formalization of Observable Information

A *σ-algebra* $\mathcal{F}$ on Ω is a collection of subsets of Ω that represents the *observable information*, i.e., all events that can be assigned a probability. Formally, $\mathcal{F} \subseteq 2^{\Omega}$ must satisfy:

(i) $\Omega \in \mathcal{F}$ (the sample space itself is measurable);
(ii) If $A \in \mathcal{F}$, then its complement $A^c = \Omega \setminus A$ is also in $\mathcal{F}$;
(iii) If $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcup_{k=1}^{\infty} A_k \in \mathcal{F}$.

The pair $(\Omega, \mathcal{F})$ is called a *measurable space*.

3.2 Probability Measure

A *probability measure* is a mapping $\mathbb{P} : \mathcal{F} \to [0, 1]$ such that:

(a) $\mathbb{P}(\Omega) = 1$;
(b) For any countable sequence of pairwise disjoint events $A_k \in \mathcal{F}$,

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mathbb{P}(A_k).$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ thus forms a *probability space*.

3.3 Random Variables and Bias as a Sub-σ-Algebra

A *random variable* is a measurable function $X : \Omega \to \mathbb{R}^d$ for some dimension $d \geq 1$. A function is measurable if for every Borel set $B \subseteq \mathbb{R}^d$ (the smallest σ-algebra generated by all open sets in $\mathbb{R}^d$), its preimage $X^{-1}(B)$ is an event in $\mathcal{F}$. Within our framework, bias is conceptualized as a structural property of a model's learned information. This information is encoded in a sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$, which is the smallest σ-algebra containing all events related to sensitive attributes (e.g., race, gender, ZIP code) and their proxies. Intuitively, $\mathcal{G}$ captures all information a model could infer about protected characteristics.

3.4 Independence from a σ-Algebra for Fairness

A random variable Y (e.g., a prediction or representation) is said to be *independent of a σ-algebra $\mathcal{G}$* if and only if, for every Borel set $B \subseteq \mathbb{R}^d$ and every $G \in \mathcal{G}$,

$$\mathbb{P}(Y \in B \mid G) = \mathbb{P}(Y \in B).$$

This definition is a direct mathematical embodiment of fairness: conditioning on any information in $\mathcal{G}$ does not alter the probability distribution of Y. This requirement is significantly stronger than zero correlation, as it mandates that the model reveals *no information* about protected attributes.

3.5 Best L^2-Approximation Independent of $\mathcal{G}$

Given a potentially biased random variable X, the goal is to construct a fair representation Y that is independent of $\mathcal{G}$ and optimally approximates X. The notion of optimality is formalized as minimizing the mean squared error in the Hilbert space $L^2(\Omega, \mathcal{F}, \mathbb{P})$. Formally, we seek $Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ such that:

(1) Y is independent of $\mathcal{G}$;

(2) Y minimizes

$$\mathbb{E}\big[\|X - Y\|^2\big] = \inf_{Z \perp\!\!\!\perp \mathcal{G}} \mathbb{E}\big[\|X - Z\|^2\big].$$

As established in [4,5,3], a unique solution Y exists under the mild condition that $(\Omega, \mathcal{F}, \mathbb{P})$ is *atomless conditionally on* $\mathcal{G}$. The construction leverages optimal transport theory, solving for the optimal distribution and subsequently realizing it as a random variable independent of $\mathcal{G}$.

3.6 A Case Study: Algorithmic Bias in Loan Approval

To ground our theoretical framework in a socially consequential domain [12,1,13], we consider the problem of *loan approval*. Credit scoring and underwriting systems directly influence access to financial resources and opportunities, yet numerous studies and legal cases have revealed systematic disparities in approval rates across gender, racial, and socioeconomic lines. Such disparities often arise not because sensitive attributes are explicitly included in the model, but because correlated features act as *proxies* for protected characteristics. This makes loan approval a paradigmatic example of algorithmic bias, and a suitable case study for the framework we develop in this work.

Problem Formulation in a Probability Space We formalize the loan approval process as a probability space $(\Omega, \mathcal{F}, \mathbb{P})$:

- The **sample space** Ω is the set of all potential loan applicants. Each element $\omega \in \Omega$ corresponds to one applicant with a complete set of characteristics.
- The σ**-algebra** $\mathcal{F}$ is the collection of events measurable by the underwriting process. These events encode information about an applicant's credit history, income, debt-to-income ratio, employment record, and other relevant features.
- The **probability measure** $\mathbb{P}$ encodes the distribution of applicants in the population. In practice, $\mathbb{P}$ corresponds to the empirical distribution over a historical dataset of loan applicants.

Random Variables and the σ-Algebra of Bias Within this probability space, the following random variables capture the structure of the problem:

- $X : \Omega \to \mathbb{R}^d$ denotes the **original feature vector** available to the lender's model, consisting of variables such as income, credit score, debt-to-income ratio, and employment length.
- $S : \Omega \to \{\text{male}, \text{female}\}$ represents a **sensitive attribute**, here gender. The information content of S generates a sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$ that includes not only gender itself, but also any measurable event that acts as a proxy (e.g., certain educational backgrounds, given names, or residential histories).
- Y is the **fair representation** of an applicant's financial profile, constructed to be independent of $\mathcal{G}$ while remaining the closest possible approximation to X in the L^2 sense.

The central challenge is that even if the lender does not explicitly include S in X, correlated variables may allow the model to reconstruct gender information, leading to discriminatory outcomes. By requiring $Y \perp\!\!\!\perp \mathcal{G}$, our framework guarantees that no event in $\mathcal{G}$—including both S and its proxies—can influence the distribution of Y. This independence criterion enforces a stringent mathematical notion of fairness, ensuring that decisions derived from Y cannot be biased by sensitive information.

In subsequent sections, we will leverage *ontologies* to systematize the identification of proxies for sensitive attributes, thereby providing a principled method to specify the sub-σ-algebra $\mathcal{G}$ in complex domains. This case study will serve as a running example throughout the paper.

4 Ontologies: A Formal Foundation for Knowledge Representation

In our framework, we employ *ontologies* to provide a formal and explicit specification of domain knowledge. An ontology is a logical theory that formally defines a shared conceptualization of a domain [8,9]. Unlike a simple data schema, an ontology provides a rich set of logical axioms that enable automated reasoning and inference. We adopt a view of ontologies grounded in *Description Logic (DL)*, a decidable fragment of first-order logic widely used in knowledge representation.

4.1 Formal Components

An ontology $\mathcal{O}$ is a tuple

$$\mathcal{O} = (\mathcal{C}, \mathcal{R}, \mathcal{I}, \mathcal{A}),$$

where:

- $\mathcal{C}$ is a finite set of *class symbols* (concepts);
- $\mathcal{R}$ is a finite set of *role symbols* (binary relations or properties);
- $\mathcal{I}$ is a finite set of *individual symbols*;
- $\mathcal{A}$ is a finite set of *axioms*.

The meaning of these symbols is defined by an *interpretation*

$$\mathcal{M} = (\Delta^{\mathcal{M}}, \cdot^{\mathcal{M}}),$$

where $\Delta^{\mathcal{M}}$ is a nonempty set called the *domain of interpretation*, and $\cdot^{\mathcal{M}}$ is an interpretation function mapping symbols to elements and relations in this domain:

- Each individual $a \in \mathcal{I}$ is mapped to an element $a^{\mathcal{M}} \in \Delta^{\mathcal{M}}$;
- Each class $C \in \mathcal{C}$ is mapped to a subset $C^{\mathcal{M}} \subseteq \Delta^{\mathcal{M}}$;
- Each role $r \in \mathcal{R}$ is mapped to a binary relation $r^{\mathcal{M}} \subseteq \Delta^{\mathcal{M}} \times \Delta^{\mathcal{M}}$.

In our setting, the domain $\Delta^{\mathcal{M}}$ corresponds directly to the sample space Ω from the probability space formulation. Thus, we can formally state the equivalence:

$$\Delta^{\mathcal{M}} \equiv \Omega.$$

4.2 Knowledge Base as an Axiomatic System

An ontology's knowledge is encapsulated in a set of logical axioms $\mathcal{A}$, typically partitioned into two components: the *TBox* and the *ABox*.

- The **TBox** (Terminological Box), denoted $\mathcal{T}$, contains axioms that define relationships between classes and roles. These are general statements about the domain. For example, a subsumption axiom $C \sqsubseteq D$ is satisfied if $C^{\mathcal{M}} \subseteq D^{\mathcal{M}}$ for every interpretation $\mathcal{M}$, meaning that every instance of class C is also an instance of class D.
- The **ABox** (Assertional Box), denoted $\mathcal{A}$, contains axioms about specific individuals. These are factual statements about the domain. For example, a class assertion $C(a)$ is satisfied if $a^{\mathcal{M}} \in C^{\mathcal{M}}$, and a role assertion $r(a, b)$ is satisfied if $(a^{\mathcal{M}}, b^{\mathcal{M}}) \in r^{\mathcal{M}}$.

Reasoning over an ontology involves determining whether a given axiom is a logical consequence of the knowledge base. This is denoted

$$\mathcal{O} \models \alpha,$$

meaning that the axiom α is true in all interpretations that satisfy the axioms in $\mathcal{O}$.

4.3 A Case Study: Ontological Modeling of Loan Approval Bias

To illustrate how ontologies provide a rigorous foundation for capturing algorithmic bias, we revisit the loan approval problem using the formal components of an ontology. Recall that an ontology is defined as a tuple

$$O = (C, R, I, A),$$

where C is a set of classes, R a set of roles, I a set of individuals, and A a set of axioms.

- **Individuals** (I): Each loan applicant corresponds to an individual symbol. For example, $i_1 = \texttt{JohnDoe}$ and $i_2 = \texttt{JaneSmith}$.
- **Classes** (C): We define domain-relevant classes such as

  $\texttt{LoanApplicant}$, $\texttt{HighRiskApplicant}$, $\texttt{ApprovedApplicant}$, $\texttt{ProxyForLowIncome}$.

- **Roles** (R): These encode binary relations, for example:

  $\texttt{hasCreditScore}$, $\texttt{livesInZIP}$, $\texttt{hasEmploymentStatus}$.

The semantics are given by an interpretation $M = (\Delta^M, \cdot^M)$, where the domain Δ^M corresponds to the set of all loan applicants (i.e., $\Delta^M \equiv \Omega$ from the probability space formulation). The interpretation function $\cdot^M$ maps:

- an individual symbol $a \in I$ to an element $a^M \in \Delta^M$,
- a class $C \in C$ to a subset $C^M \subseteq \Delta^M$,
- a role $r \in R$ to a binary relation $r^M \subseteq \Delta^M \times \Delta^M$.

The knowledge base consists of axioms $A = T \cup A'$, where T is the TBox and A' is the ABox:

- **TBox axioms** define structural relationships. For instance:

  $$\exists \texttt{livesInZIP}.\{\texttt{ZIP\_12345}\} \sqsubseteq \texttt{ProxyForLowIncome}, \texttt{ProxyForLowIncome} \sqsubseteq \texttt{SensitiveAttribute}.$$

  This encodes that applicants living in ZIP code 12345 belong to a proxy class for low income, which is classified as sensitive.
- **ABox axioms** describe assertions about individuals. For example:

  $$\texttt{LoanApplicant}(\texttt{JohnDoe}), \texttt{hasCreditScore}(\texttt{JohnDoe}, 580), \texttt{livesInZIP}(\texttt{JohnDoe}, \texttt{ZIP\_12345}).$$

  Reasoning over this ontology reveals that:

  $$\texttt{JohnDoe} \in (\texttt{ProxyForLowIncome})^M,$$

which means that the applicant is inferred to be associated with a sensitive proxy. If this information is used in loan decision-making, the system can be flagged as exhibiting bias.

Thus, the ontological framework provides a precise, formal mechanism for :

1. identifying proxies of sensitive attributes,
2. enforcing fairness constraints via axioms, and
3. ensuring that the reasoning process remains transparent and explainable.

5 Ontology-Guided σ-Algebras

The central challenge in bias formalization is identifying the precise set of events that encode sensitive attributes and their proxies. In a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, fairness requires that a model's output Y be independent of a sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$ that captures all "unsettling information." The difficulty lies in systematically and completely specifying $\mathcal{G}$: naive approaches that manually list features are ad hoc and prone to omitting subtle proxies.

5.1 Integrating Ontologies and σ-Algebras

The central challenge in bias formalization is identifying the precise set of events that encode sensitive attributes and their proxies. In our setting, the sample space Ω is fixed as the finite set of data subjects (e.g., loan applicants or database rows). We equip Ω with the discrete σ-algebra $F = 2^{\Omega}$ and a probability measure P (e.g., the empirical distribution over rows). For each sensitive or proxy concept C, entailment in the ontology allows us to identify the subset of Ω that falls under C. This guarantees that all such events are measurable, since F contains all subsets of Ω.

Definition 1 (Ontology-Guided σ-Algebra of Bias). *Let $O = (C, R, I, A)$ be an OWL 2-QL ontology with TBox T and ABox A. Let $S_{onto} \subseteq C$ contain every concept name explicitly declared as a sensitive or proxy attribute in O.*

Fix a countable sample space Ω (e.g., the finite set of data-subject identifiers or database rows). Equip Ω with the discrete σ-algebra $F = 2^{\Omega}$ and a probability measure P. For each concept $C \in S_{onto}$ define its extension

$$[\![C]\!] = \{\omega \in \Omega \mid O \models C(\omega)\},$$

where the entailment relation $\models$ is evaluated in OWL 2-QL (i.e., by rewriting into a SQL query over the ABox).

The ontology-guided σ-algebra of bias *is*

$$\mathcal{G}_O = \sigma\big(\{[\![C]\!] \mid C \in S_{onto}\}\big) \subseteq F.$$

Because Ω is finite, $\mathcal{G}_O$ is finite, uniquely determined by the ABox, and computable in AC_0 data complexity.

Intuitively, $\mathcal{G}_O$ is the smallest σ-algebra on Ω that makes the class extensions of all sensitive and proxy concepts measurable. Because Ω is finite, the construction is unique and computable by standard OWL 2-QL query answering. This shifts the fairness specification problem from ad hoc feature selection to a principled and auditable pipeline: sensitive attributes are declared in the ontology, compiled into measurable events via entailment, and collected into the bias σ-algebra.

Why Ontology-Guided σ-Algebras are Robust. This integration offers three decisive advantages:

1. **Completeness.** Ontological axioms in the TBox allow systematic discovery of proxy attributes through logical inference. For example, an axiom such as

$$\mathsf{livesInZIP}(x, y) \wedge \mathsf{Redlined}(y) \;\Rightarrow\; \mathsf{ProxyForRace}(x)$$

   ensures that $\mathcal{G}_O$ includes not only explicit sensitive attributes but also indirect dependencies like ZIP $\rightarrow$ race, even when such correlations are not obvious in the raw data.
2. **Transparency and Auditability.** The TBox serves as a machine-executable fairness policy. Regulators or auditors can inspect axioms to verify which attributes and proxies generate measurable events $[\![C]\!]$, making the fairness criterion both human-readable and certifiable.
3. **Mathematical Rigor.** By defining $\mathcal{G}_O$ as a true σ-algebra generated by ontology-specified events $[\![C]\!]$, subsequent fairness operations (e.g., constructing $Y \perp\!\!\!\perp \mathcal{G}_O$) are applied to a formally closed and comprehensive collection of events. This removes arbitrariness in feature selection and guarantees provable independence.

The ontology-guided σ-algebra $\mathcal{G}_O$ therefore transforms the informal task of "finding sensitive proxies" into a principled construction grounded in logic and measure theory. It provides a complete, transparent, and mathematically robust basis for certifiable fairness, bridging symbolic knowledge representation and probabilistic independence.

5.2 Why Ontology-Guided σ-Algebras are Robust

This integration offers three decisive advantages:

1. **Completeness:** Ontological axioms in $\mathcal{T}$ allow systematic discovery of proxy attributes through logical inference. For example, axioms of the form

$$\texttt{livesInZIP}(x, y) \wedge \texttt{Redlined}(y) \Rightarrow \texttt{ProxyForRace}(x)$$

   ensure that $\mathcal{G}_\mathcal{O}$ captures indirect dependencies (ZIP code $\rightarrow$ race) that may not be evident through correlation analysis alone.
2. **Transparency and Auditability:** The TBox serves as a machine-executable fairness policy. Regulators or auditors can directly inspect axioms to verify which attributes and proxies generate $\mathcal{G}_\mathcal{O}$, making the fairness criterion both human-readable and certifiable.
3. **Mathematical Rigor:** By defining $\mathcal{G}_\mathcal{O}$ as a true σ-algebra generated by ontology-specified events, subsequent fairness operations (e.g., constructing $Y \perp\!\!\!\perp \mathcal{G}_\mathcal{O}$) are applied to a formally closed and comprehensive collection of events. This eliminates the arbitrariness of feature selection and guarantees provable independence.

The ontology-guided σ-algebra $\mathcal{G}_{\mathcal{O}}$ transforms the informal task of "finding sensitive proxies" into a principled construction grounded in logic and measure theory. It provides a complete, transparent, and mathematically robust basis for certifiable fairness, bridging symbolic knowledge representation and probabilistic independence.

5.3 A Case Study: Ontology-Guided σ-Algebra in Loan Approval

To demonstrate the robustness of our framework, we revisit the loan approval problem, integrating the measure-theoretic and ontological perspectives. Traditional debiasing methods often fail because they treat bias superficially— removing a small set of correlated features while missing subtle proxies. In contrast, our ontology-guided σ-algebra $\mathcal{G}_O$ formalizes and exhaustively captures all sensitive and proxy information, ensuring rigorous independence.

Scenario: The Proxy Problem. Suppose a bank's loan approval system is trained without explicit sensitive variables such as gender or race, but uses features including an applicant's residential ZIP code and prior educational institution. While neither feature is directly sensitive, both are well-known proxies:

- Certain ZIP codes in the United States are historically linked to racial segregation or socio-economic disadvantage.
- Educational institutions, such as historically Black colleges or single-gender institutions, carry implicit demographic information about applicants.

Naive debiasing methods that simply drop the ZIP code feature are insufficient: the model may still exploit correlations through the educational institution or other variables. The bias "flows" through alternative proxies.

Ontology-Guided Specification. We construct an ontology $O = (C, R, I, A)$ to formally encode this domain knowledge:

- Classes (C): LoanApplicant, UrbanArea, SuburbanArea, HistoricallyBlackCollege, StateUniversity, ProxyForRace, ProxyForSES.
- Roles (R): livesIn (applicant $\to$ area), attended (applicant $\to$ institution), hasCreditScore (applicant $\to \mathbb{R}$).
- Individuals (I): specific applicants such as JohnDoe, JaneSmith.
- Axioms (A): logical rules linking classes and roles to sensitive attributes. For example:

$$\mathsf{attended}(x, y) \wedge y \in \mathsf{HistoricallyBlackCollege} \;\Rightarrow\; x \in \mathsf{ProxyForRace},$$

$$\mathsf{livesIn}(x, z) \wedge z \in \mathsf{UrbanArea} \wedge \mathsf{MedianIncome}(z) < 30{,}000 \;\Rightarrow\; x \in \mathsf{ProxyForSES}.$$

Ontology-Guided σ-Algebra. For each sensitive or proxy concept $C \in S_{\text{onto}}$, we compute its extension

$$[\![C]\!] = \{\omega \in \Omega \mid O \models C(\omega)\}.$$

The ontology-guided σ-algebra of bias is then

$$\mathcal{G}_O = \sigma\big(\{[\![C]\!] \mid C \in S_{\text{onto}}\}\big).$$

In this example, $\mathcal{G}_O$ includes events such as:

- the set of applicants living in ZIP codes designated as socio-economic proxies,
- the set of applicants who attended institutions correlated with race or gender,
- all Boolean combinations of such events.

Robustness of the Approach. This ontology-guided construction ensures:

1. **Completeness:** All sensitive and proxy variables entailed by the ontology are automatically included in $\mathcal{G}_O$. This guards against residual bias hidden in indirect correlations.
2. **Transparency:** The TBox axioms explicitly document the fairness policy, making $\mathcal{G}_O$ human-readable and auditable.
3. **Provable Fairness:** By constructing a fair representation $Y \in L^2(\Omega, F, P)$ such that

$$Y \perp\!\!\!\perp \mathcal{G}_O, \quad \mathbb{E}\big[\|X - Y\|^2\big] = \inf_{Z \perp\!\!\!\perp \mathcal{G}_O} \mathbb{E}\big[\|X - Z\|^2\big],$$

We obtain certifiable independence from all sensitive and proxy information.

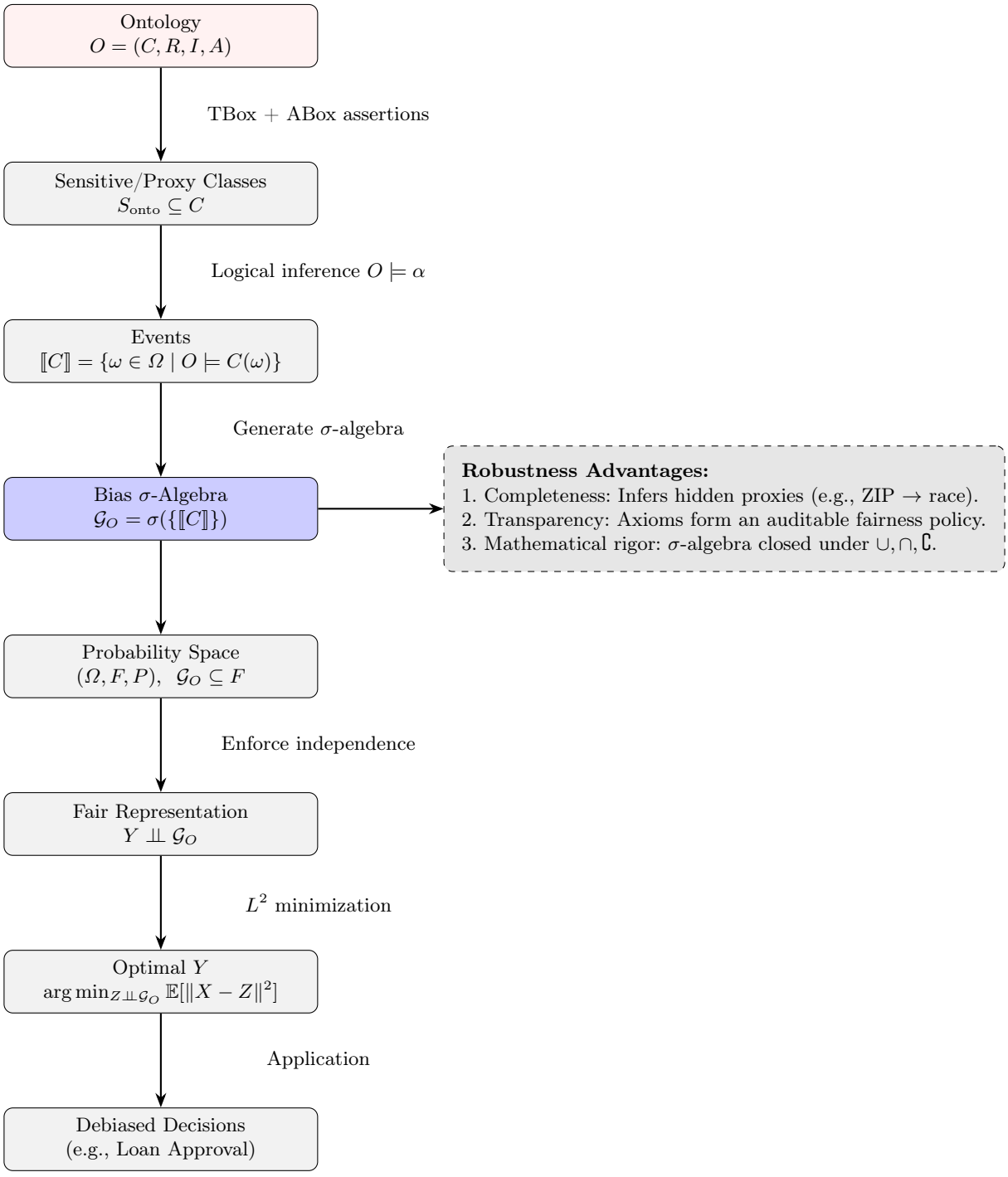

Fig. 1. Ontology-guided σ-algebra construction and robustness in AI fairness. Sensitive and proxy classes S_{onto} are identified in the ontology, extended to measurable events $[\![C]\!]$, and generate the bias σ-algebra $\mathcal{G}_O \subseteq F$. A fair representation Y is then constructed such that $Y \perp\!\!\!\perp \mathcal{G}_O$ and is optimal in L^2, ensuring debiased loan approval decisions while guaranteeing completeness, transparency, and mathematical rigor.

The ontology-guided σ-algebra $\mathcal{G}_O$ thus provides a rigorous, auditable, and mathematically grounded way to capture the complete structure of bias in loan approval, ensuring that fairness is enforced not only for explicitly sensitive variables but also for their hidden proxies.

6 Implementation Blueprint

This section outlines a prospective engineering plan that translates the theoretical framework of Sections §3, §4 and §5 into a reproducible pipeline. The objective is to enable practitioners to move from *ethical intent* (encoded in an OWL 2-QL ontology) to *provably fair decisions* with minimal friction. At this stage, the blueprint remains conceptual and intended as a roadmap for future implementation, leveraging established tools like the open-source OWL reasoner HermiT [7] and benchmark fairness datasets like COMPAS [6] to ensure practical grounding.

6.1 End-to-End Pipeline

Figure 2 illustrates the four macro-stages that form our reference implementation design:

1. **Ontology Authoring** (§6.2)
2. **σ-Algebra Generator** (§6.3)
3. **Fair Representation Engine** (§6.4)
4. **Certification & Audit** (§6.5)

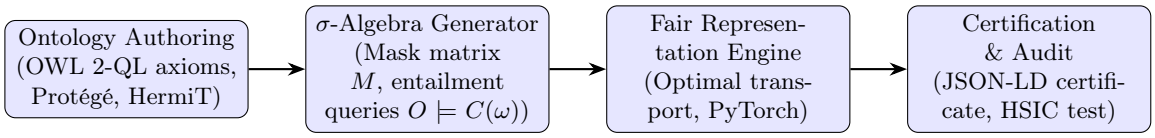

Fig. 2. Planned implementation pipeline: sensitive and proxy classes S_{onto} are declared in the ontology, compiled into measurable events $[\![C]\!]$, and aggregated into the bias σ-algebra $\mathcal{G}_O$. Optimal transport then constructs a fair representation $Y \perp\!\!\!\perp \mathcal{G}_O$, and certification provides auditability.

6.2 Ontology Authoring Guidelines

We propose structuring the ontology in three layers:

- **Upper Layer:** Generic fairness vocabulary (e.g., `SensitiveAttribute`, `ProxyAttribute`).
- **Domain Layer:** Mid-level domain concepts such as `LoanApplicant`, `ZIPCode`, `HBCU`.

– **Policy Layer:** Project-specific axioms, e.g.,

$$\texttt{livesIn}(x, \texttt{zip}) \wedge (\texttt{MedianIncome}(\texttt{zip}) < 30{,}000) \longrightarrow \texttt{ProxyForSES}(x).$$

The final ontology is serialized in RDF/XML and passed as input to the σ-Algebra Generator.

6.3 σ-Algebra Generator

Inputs: (1) OWL 2-QL ontology file, (2) dataset $\mathcal{D} = \{\omega_i\}_{i=1}^{N}$, (3) mapping from ontology individuals to dataset rows.

Output: A binary mask matrix $M \in \{0,1\}^{N \times k}$ where

$$M_{i,j} = 1 \iff O \models C_j(\omega_i), \quad C_j \in S_{\text{onto}}.$$

Each column corresponds to an event $[\![C_j]\!]$, and the set $\{[\![C_j]\!] \mid C_j \in S_{\text{onto}}\}$ generates the bias σ-algebra $\mathcal{G}_O$.

Complexity: Computing M reduces to $|S_{\text{onto}}|$ SPARQL `ASK` queries per row, i.e., $\mathcal{O}(N \cdot |S_{\text{onto}}|)$. With columnar storage (Parquet/ORC) and vectorized query engines, datasets with up to 10^7 rows should be processable on a single multi-core node within practical time limits.

6.4 Fair Representation Engine

We plan to implement the Delbaen–Majumdar optimal-transport solver in PyTorch. Given M and the original feature matrix $X \in \mathbb{R}^{N \times d}$, the procedure will:

1. Estimate conditional expectations $\hat{\mu}_g = \mathbb{E}[X \mid g]$ for each g in the partition induced by M (intersections of $[\![C_j]\!]$).
2. Solve the discrete optimal-transport problem to obtain coupling $\pi^\star$.
3. Map X to Y such that $Y \perp\!\!\!\perp \mathcal{G}_O$.

Algorithm 1 Planned Optimal-Transport Fair Projection

Require: $X \in \mathbb{R}^{N \times d}$, mask matrix M, regularization $\varepsilon > 0$
1: Build partition $\{\Omega_g\}_{g \in \mathcal{G}_O}$ from unique rows of M
2: Compute $\mu_g = \frac{1}{|\Omega_g|} \sum_{i \in \Omega_g} X_i$ for all g
3: Solve entropic OT: $\pi^\star = \arg\min_\pi \sum_{i,j} \pi_{ij} \|X_i - \mu_{g_j}\|^2 + \varepsilon H(\pi)$
4: Return $Y_i = \sum_j \pi^\star_{ij}\, \mu_{g_j}$

Complexity: Naively $\mathcal{O}(N^2 d)$, but Sinkhorn iterations on GPU are expected to reduce wall-clock time to seconds for $N \approx 10^6$.

6.5 Certification & Audit

Planned Certificate Contents: Each model would be accompanied by a JSON-LD certificate containing:

- SHA-256 hash of the ontology file,
- hash of the mask matrix M,
- reconstruction error $\|X - Y\|_F^2/N$,
- independence test p-value via HSIC [17].

Regulatory Audit Hook: Third-party auditors could re-run the pipeline with the disclosed ontology and dataset (e.g., COMPAS [6]) to verify that the generated σ-algebra $\mathcal{G}_O$ matches the training-time one.

6.6 Deployment Footprint

We envisage packaging the system as:

- A `docker-compose` stack (OWL reasoner like HermiT, Spark job for M, PyTorch service for OT).
- Helm charts for Kubernetes with autoscaling GPU nodes.
- A minimal CPU-only fallback for development environments.

This modular design balances scalability with accessibility. Performance estimates (e.g., end-to-end latency for multi-million-row datasets) remain to be validated in a full implementation.

7 Conclusion

This work presents a novel framework that bridges the gap between ethical declarations and mathematical guarantees in AI fairness by integrating ontology engineering with measure-theoretic optimal transport. Our approach addresses a fundamental limitation in existing bias mitigation techniques: the inability to systematically identify and eliminate all forms of sensitive information, including subtle proxies that enable discriminatory decision-making through seemingly neutral features.

The theoretical foundation of our framework rests on three key innovations. First, we formalize algorithmic bias as a structural property of a model's learned σ-algebra, providing a mathematically rigorous characterization that extends beyond simple correlation-based measures. Second, we demonstrate how OWL 2-QL ontologies can be systematically compiled into σ-algebras by defining class extensions $[\![C]\!] = \{\omega \in \Omega \mid O \models C(\omega)\}$ for sensitive and proxy concepts. This construction ensures that $\mathcal{G}_O = \sigma(\{[\![C]\!]\})$ captures both explicitly declared sensitive attributes and their proxies derived by logical inference. Third, we establish that Delbaen–Majumdar optimal transport provides the unique solution for constructing fair representations that are provably independent of biased information while minimizing information loss in the L^2-norm.

Our loan approval case study illustrates the practical significance of this approach. Traditional debiasing methods that merely remove explicitly sensitive features fail to address the "proxy problem," where correlated variables such as ZIP codes or educational institutions continue to encode protected characteristics. By contrast, our ontology-guided σ-algebra $\mathcal{G}_O$ systematically identifies these indirect pathways to bias, ensuring that the resulting fair representation Y satisfies $Y \perp\!\!\!\perp \mathcal{G}_O$—a guarantee of exact statistical independence rather than mere decorrelation.

The framework's robustness stems from three critical advantages that distinguish it from existing approaches. *Completeness* is achieved through ontological reasoning that systematically discovers proxy relationships, preventing residual bias from hidden correlations. *Transparency* emerges from the explicit axiomatization of fairness policies in machine-readable form, enabling regulatory audit and public scrutiny. *Mathematical rigor* is ensured through the closure properties of σ-algebras and optimal transport, providing certifiable guarantees rather than heuristic approximations.

While we have established the theoretical foundations and provided a detailed implementation blueprint, several important directions warrant future investigation. Empirical validation across diverse domains and datasets remains essential to demonstrate the framework's practical effectiveness and computational scalability. The proposed pipeline, leveraging OWL 2-QL reasoning and PyTorch-based optimal transport solvers, requires full implementation and performance evaluation on large-scale datasets. Additionally, the framework's behavior under distribution shift and its integration with existing MLOps pipelines merit careful study.

The broader implications of this work extend beyond technical considerations to fundamental questions about the role of AI in society. As AI systems increasingly influence consequential decisions in lending, hiring, healthcare, and criminal justice, the need for mathematically grounded fairness guarantees becomes paramount. Our framework provides a pathway from informal ethical commitments to formal mathematical constraints, offering a foundation for building AI systems that are not merely performant but demonstrably equitable.

By ensuring that fairness constraints are both comprehensive and mathematically verifiable, the integration of symbolic reasoning through ontologies with probabilistic independence through optimal transport contributes to the broader goal of developing trustworthy AI systems that serve all members of society equitably.

References

1. Adegoke, T., Ofodile, O., Ochuba, N., Akinrinol, O.: Evaluating the fairness of credit scoring models: A literature review on mortgage accessibility for underreserved populations. GSC Advanced Research and Reviews **18**(3), 189–199 (2024)
2. Davis, S.E., Dorn, C., Park, D.J., Matheny, M.E.: Emerging algorithmic bias: fairness drift as the next dimension of model maintenance and sustainability. Journal of the American Medical Informatics Association **32**(5), 845–854 (2025)

3. Delbaen, F., Majumdar, C.: Approximation with independent variables. In: Peter Carr Gedenkschrift Research Advances in Mathematical Finance, chap. 9, pp. 311–327. World Scientific Publishing Co. Pte. Ltd. (2023), https://EconPapers.repec.org/RePEc:wsi:wschap:9789811280306_0009
4. Delbaen, F., Majumdar, C.: Approximation with independent variables. In: Peter Carr Gedenkschrift: Research Advances in Mathematical Finance, pp. 311–327. World Scientific (2024)
5. Delbaen, F., Majumdar, C.: Convergence results for approximation with independent variables. arXiv preprint arXiv:2405.19780 (2024)
6. Fabris, A., Messina, S., Silvello, G., Susto, G.A.: Algorithmic fairness datasets: the story so far. Data Mining and Knowledge Discovery **36**(6), 2074–2152 (2022)
7. Glimm, B., Horrocks, I., Motik, B., Stoilos, G., Wang, Z.: Hermit: an owl 2 reasoner. Journal of automated reasoning **53**(3), 245–269 (2014)
8. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge acquisition **5**(2), 199–220 (1993)
9. Horrocks, I.: Description logic: A formal foundation for ontology languages and tools. Methods In Cell Biology **78**, 765–775 (2007)
10. Jain, S., Hamidieh, K., Georgiev, K., Ilyas, A., Ghassemi, M., Madry, A.: Data debiasing with datamodels (d3m): Improving subgroup robustness via data selection. arXiv preprint arXiv:2406.16846 (2024)
11. Joachimiak, M.P., Miller, M.A., Caufield, J.H., Ly, R., Harris, N.L., Tritt, A., Mungall, C.J., Bouchard, K.E.: The artificial intelligence ontology: Llm-assisted construction of ai concept hierarchies. Applied Ontology **19**(4), 408–418 (2024)
12. Kelly, S., Mirpourian, M.: Algorithmic bias, financial inclusion, and gender. Women's World Banking (2021)
13. Ongena, S., Popov, A.: Gender bias and credit access. Journal of Money, Credit and Banking **48**(8), 1691–1724 (2016)
14. Safranek, S., Zvácková, A.: Ontological framework for integrating predictive analytics, ai, and big data in decision-making systems using knowledge graph. In: Yang, X., Drogoul, A., Wagner, G. (eds.) Proceedings of the 15th International Conference on Simulation and Modeling Methodologies, Technologies and Applications, SIMULTECH 2025, Bilbao, Spain, June 11-13, 2025. pp. 287–294. SCITEPRESS (2025). https://doi.org/10.5220/0013514400003970, https://doi.org/10.5220/0013514400003970
15. Scandizzo, S., Majumdar, C.: Escaping the geometry of bias via independence from learned o-algebras. Springer (2025)
16. Wang, A., Phan, M., Ho, D.E., Koyejo, S.: Fairness through difference awareness: Measuring desired group discrimination in llms (2025), https://arxiv.org/abs/2502.01926
17. Wang, T., Dai, X., Liu, Y.: Learning with hilbert–schmidt independence criterion: A review and new perspectives. Knowledge-based systems **234**, 107567 (2021)
18. Zhou, W., Liu, F., Zheng, H., Zhao, R.: Mitigating data bias and ensuring reliable evaluation of ai models with shortcut hull learning. Nature Communications **16**(1), 5513 (2025)