Towards an Ontology-Driven Approach to Document Bias

MAYRA RUSSO*, L3S Research Center & Leibniz University Hannover, Germany
MARIA-ESTHER VIDAL, TIB Leibniz Information Center for Science and Technology, L3S Research Center & Leibniz University of Hannover, Germany

Machine learning (ML)-powered systems are capable of reproducing and often amplifying undesired biases embedded in society, emphasizing the importance of operating under practices that enable the study and understanding of the intrinsic characteristics of ML pipelines. This supports the emergence of documentation frameworks with the idea that "any remedy for bias starts with awareness of its existence." However, a resource that can formally describe ML pipelines in terms of detected biases is still missing. To address this gap, we present the Doc-BiasO ontology, a resource that sets out to create an integrated vocabulary of biases defined in the *Trustworthy AI* literature and their measures, as well as to incorporate relevant domain terminology and relationships between them. Overseeing ontology engineering best practices, we reuse existing vocabularies on machine learning and AI to foster knowledge sharing and interoperability between the actors concerned with its research, development, regulation, and others. In addition, we demonstrate the potential of Doc-BiasO with an experiment on an existing benchmark and as part of a neuro-symbolic system. Overall, our main objective is to contribute towards clarifying existing terminology on bias research as it rapidly expands to all areas of AI and to improve the interpretation of bias in data and downstream impact through its documentation.

JAIR Track: Fairness and Bias in AI

JAIR Associate Editor: Roberta Calegari

JAIR Reference Format:
Mayra Russo and Maria-Esther Vidal. 2025. Towards an Ontology-Driven Approach to Document Bias. *Journal of Artificial Intelligence Research* 83, Article 38 (August 2025), 35 pages. DOI: 10.1613/jair.1.19388

1 Introduction

The breakthroughs and benefits attributed to machine learning (ML)-powered systems, or *AI* in broader terms, prompted in great part by the abundance of available data [79, 67], have also helped to make prevalent how these systems are capable of producing unexpected, biased, and in some cases undesirable output [12, 11]. Some examples of seminal work on harmful biases (i.e., a concentration on or interest in one particular area or subject, often considered to be unfair) in the context of AI systems deployed in various applications across different domains have demonstrated how facial recognition tools and popular search engines can exacerbate demographic disparities, worsening the marginalization of minorities at the individual and group level [18, 52]. Other examples of research on this topic have demonstrated how biases in news recommenders and social media feeds actively play a role in conditioning and manipulating people's behavior and amplifying individual and public opinion polarization [8, 9].

*Corresponding Author.

Authors' Contact Information: Mayra Russo, ORCID: 0000-0001-7080-6331, mrusso@l3s.de, L3S Research Center & Leibniz University Hannover, Hannover, Germany; Maria-Esther Vidal, ORCID: 0000-0003-1160-8727, maria.vidal@tib.eu, TIB Leibniz Information Center for Science and Technology, L3S Research Center & Leibniz University of Hannover, Hannover, Germany.

DOI: 10.1613/jair.1.19388

The notoriety of bias in relation to AI systems that resort to machine learning (ML) methods has thus caused the introduction of multiple measures to discover, account for, and mitigate its detrimental impact. Further, it has also given a platform for scholars to advocate for practices that account for the characteristics of ML pipelines (i.e., datasets, ML models, and user interaction loop) [48] to enable actors concerned with its research, development, regulation, and use to inspect all actions performed throughout the engineering process, with the objectives to account for and mitigate bias, as well as to increase trust placed not only on the development processes but on the systems themselves [28, 61, 77, 60].

The AI community does not have standardized methodologies, neither to measure biases nor to produce documentation on AI pipelines, nor are there regulatory frameworks that enforce these practices at the moment of writing. Despite this, pioneering work on human-readable (i.e., textual descriptions in a format that humans can read and understand) documentation frameworks for machine learning pipelines argues that "drawing on values-sensitive practices can only bring about improvements in engineering and scientific outcomes" [15]. Similarly, there is the argument that documentation promotes the communication between "consumers and producers" [28], while making a case for how exhaustive documentation of the characteristics of these artifacts can support the identification of biases reflected in them. To this point, semantic data models (e.g., ontologies, knowledge graphs) can also play a crucial role in supporting the implementation of bias assessments, bias representation, and bias mitigation tasks [65] in a way that is also machine-readable (i.e., makes available a fine-grained description of data in a format manageable by computers). This characteristic improves the findability, accessibility, interoperability, and reusability (*FAIR*) of data-centric resources [53, 46] and also positions them to be used in the elaboration of documentation for AI systems by enhancing their accuracy and interpretability [24, 14].

Ontologies to model existing machine learning fairness metrics [25, 26], as well as the semantic specifications to catalog risks in terms of compliance and conformance of AI systems under the EU's AI Act[1] [30, 31] have been proposed; however, a resource that can formally describe ML pipelines and provides a vocabulary to characterize them in terms of measured biases is still amiss.

Proposed Solution We propose an ontology-driven approach to describe and document biases detected across ML pipelines. Here, we refer to documentation as the process of generating metadata represented in formats understandable by humans and also by machines [53], where formal data models such as ontologies and controlled vocabularies provide standardized concepts to express this metadata. Figure 1 shows a visual summary of the main points to be addressed throughout this work. Specifically, overseeing ontology engineering best practices, our ontology, Doc-BiasO, is a resource developed with the objective of introducing an integrated vocabulary to describe machine learning pipelines (i.e., input datasets, ML models, and output) in terms of detected biases resorting to existing bias measures (metrics or indicators) defined in the literature. Additionally, we reuse existing vocabularies on ML and AI to foster knowledge sharing and interoperability between the actors concerned with AI and bias research, development, and regulation, among others. In order to demonstrate the potential of Doc-BiasO, we include two use cases.

With this work, our main objective is to contribute towards clarifying existing terminology on bias research as it rapidly expands to all areas of AI and to improve the interpretation of bias in data and downstream impact.

Contributions This paper is an extension of our previous work [69], in which we present an abbreviated overview of our resource. The novel contributions of our current work are summarized as follows:

(1) a comprehensive description of the modeled domain, as well as a description of the methodology followed to develop our vocabulary;
(2) Doc-BiasO, an integrated vocabulary system of ML-related biases and concepts. The current version of Doc-BiasO has 390 classes, 72 object properties, 243 individuals, and is publicly available.

[1] Annex III, European Council position

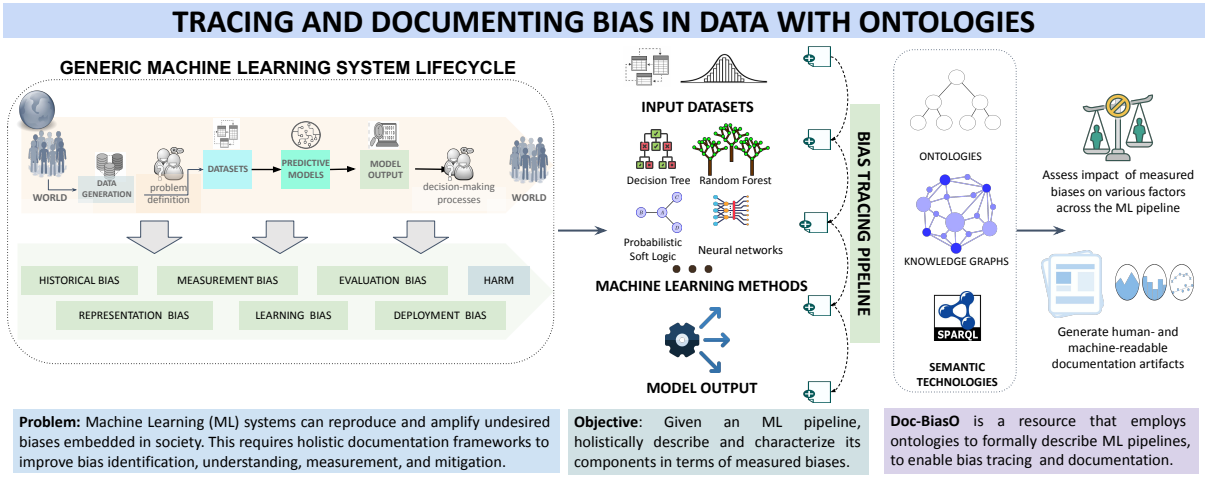


Fig. 1. **Graphical Overview.** Depicted is a visual representation of the proposed approach, which is demonstrated by first illustrating a generic machine learning system life cycle and the different types of biases that can emerge throughout it. Following this, our objective is to describe and characterize ML pipelines in terms of measured biases, resorting to existing metrics found in the literature, and encompassing input datasets to model output. Lastly, our approach proposes a comprehensive vocabulary that, by employing ontologies, can formally describe ML pipelines in order to enable bias tracing across them as well as support the generation of human- and machine-readable documentation artifacts.

(3) a finer-grained description of our modeling methodology, including the results of the alignment analysis performed for the VAIR and FMO ontologies and more details on the reused concepts from existing vocabularies and ontologies;
(4) examples of the OWL axioms that make up the Doc-BiasO ontology;
(5) a comprehensive evaluation of the coverage of represented knowledge in terms of biases modeled;
(6) an empirical demonstration of Doc-BiasO, with two examples over two generated RDF graphs applied to different domains.

The remainder of this paper is structured as follows: Section 2 presents a review of the related literature. Section 3 introduces relevant Semantic Web concepts. Section 4 describes the domain to be modeled and the details of the development of our vocabulary. While Section 5 details the design of Doc-BiasO. The results of the evaluation are reported in Section 6. Section 7 presents a practical use case of the ontology. Finally, Section 8 outlines our conclusions and future lines of work

2 Related Work

In this section, we elaborate on relevant literature, specifically on bias and machine learning, existing documentation frameworks, and existing ontologies to document machine learning pipelines.

2.1 Bias and Machine Learning

The widespread use of ML systems has shed light on the risks and ramifications associated with them in the real world. Furthermore, renowned research on the topic of algorithmic bias has demonstrated how the deployment of applications that use machine learning can exacerbate demographic disparities, worsening the marginalization of minorities at the individual and group level [18, 52].

Bias can start at any point in the ML pipeline, referring to the interconnection of data processing and modeling steps [79]. This means that not all biases emerge from data; moreover, not all biases can be measured, and even when they can, not all biases have a detrimental downstream impact. Notwithstanding, to operationalize the analysis, quantify, and understand detected biases in ML pipelines, be it at the data or model level, computational researchers primarily draw on statistical analysis and metrics to describe and characterize them.

In our work, we use bias measures (metrics and indicators) defined in the literature and incorporate them into the modeling process of our ontology. A bias measure quantitatively assesses the presence and extent of bias in a particular context. They cover the following dimensions [23]:

- *Target Group:* or entities for which bias is being assessed;
- *Attribute(s):* that may contribute to bias;
- *Group Comparison:* a method to compare across different groups based on the chosen performance metric or indicator;
- *Thresholds or Criteria:* thresholds or criteria that indicate the presence of bias.

Despite the advances in this type of technical intervention, we acknowledge and emphasize that due to the normative nature of characterizing and identifying bias [17], the definition and detection task is often highly complex, inconclusive, and unable to provide a clear-cut de-biasing "solution" [54].

2.2 Documentation Frameworks and Machine Learning

Understanding the inner workings of ML-powered systems can be hindered because of their opaqueness. Work such as - ([15, 28, 51, 37]), thus advocates for the production of value-oriented, human-readable documentation for datasets (Data Statements for Natural Language Processing, Datasheets for Datasets), ML models (Model Cards for Model Reporting), and AI systems (Use Case Cards). Doc-BiasO aims to follow their lead by combining the different components of the ML pipeline to produce comprehensive descriptions in human- and machine-readable format of data-driven pipelines.

Other documentation approaches, such as Sun et al. [78] introduce a tool to assess fitness for use of datasets. This automated data exploration tool delimits its focus to three dimensions: representativeness, bias, and correctness. In a similar vein, [82] introduces a bias visualization tool for computer vision datasets. This exploration tool narrows their assessment to three sets of metrics: object-based, gender-based, and geography-based dimensions. Further, interactive tools, developed by the industry (e.g., [63, 64, 1]), enable exploration, visualization, and comparison of datasets. The extensible and modular design of our ontology, Doc-BiasO, allows users to describe and document their data-driven pipelines and seamlessly incorporate additional descriptive dimensions and components as needed. Furthermore, the underlying knowledge-driven framework prompts the integration and fine-grained description of multiple data sources and leverages reasoning capabilities for enhanced data analytics.

2.3 Ontologies and Machine Learning

In the context of bias, the **Bias Ontology Design Pattern (BODP)** [41] is one of the first works to propose a formalization for the concept of bias. Its objective is to capture a high-level representation of bias as an abstract term and not necessarily in the context of ML systems. We reuse part of BODP as a building block; however, Doc-BiasO has a different scope and intended use.

The **fairness metrics ontology (FMO)** [25, 26] models fairness metrics (`fmo:fairness_metric`) from the literature and relates them to their use case. The conceptualization of bias and fairness in relation to ML systems is often intertwined but often does not study the same phenomena [56, 72]. Fairness in relation to ML takes the form of algorithmic interventions that incorporate mathematical formalizations of moral or legal notions for the fair treatment of different populations in ML pipelines. These interventions aim to encourage practitioners to develop ML models that satisfy the statistical non-discrimination criterion for a given subpopulation [11].

The main distinctions between FMO and Doc-BiasO are:

- The underlying framework. FMO introduces a reasoning framework to assist in the selection of fairness metrics, while we propose a descriptive vocabulary that can be used and incorporated into varying frameworks as needed. Nevertheless, Doc-BiasO is also equipped with reasoning capabilities that can be extended to further semi-automatize documentation tasks.
- The focus of our modeling is on biases in data identified in the literature and the existing measures defined to detect them. These are concepts and relations that are not made explicit in the current version of FMO.

As we consider both ontologies to be complementary, we reuse FMO to foster the development of a comprehensive vocabulary that provides coverage of terminology that pertains to the responsible development of ML systems. We follow the same approach with the **AI Risk Ontology (AIRO)** [30], and by effect, the **Vocabulary of AI Risks (VAIR)** [31]; in this case, risk in relation to ML systems, under the broader label of AI, is defined as systems that are likely to cause serious harm to the health, safety, or fundamental rights of individuals according to European Union (EU) law. These works are ontology-driven approaches to account for the compliance and conformance of AI systems under the EU's AI Act's specifications.[2] Specifically, AIRO is a modular ontology created to identify whether an AI system is classified as high-risk, while VAIR provides semantic specifications for cataloging AI risks, reusing core concepts in AIRO (e.g., `airo#Risk`, `airo#Consequence`).

Lastly, [5] proposes a descriptive framework **(ACROCPoLis)** to describe ML systems and their societal impact by making explicit the interrelations and divergent perspectives of relevant stakeholders (individuals, groups of people, and institutions). While this is beyond the scope of our work, a study would be undertaken to examine vocabulary reuse and the corresponding extension of Doc-BiasO ontology should the conceptual model be formalized and published.

The Semantic Web community has also proposed other technical solutions to improve the interpretability and transparency of machine learning pipelines. The provenance ontology (PROV-O) [44] enables the representation of provenance information generated by different entities and can be easily applied to multiple contexts. Standard schemas for data mining and machine learning algorithms, such as the Machine Learning Schema (MLS) ontology [58] and the Description of a Model (DOAM) ontology,[3] provide fine-grained vocabularies to represent the characteristics of ML models. Moreover, the question of reproducibility in ML has also been addressed [3]. Correspondingly, the Data Catalog Vocabulary (DCAT) [2] enables fine-grained descriptions of datasets and data services using a rich controlled vocabulary.

Adherent to best practices in ontology engineering [33], all these ontologies and vocabularies have been reused in the composition of Doc-BiasO.

3 Preliminaries: Semantic Data Models

Semantic data models can be defined as formal, structured, and standardized data structures, which make explicit the meaning of information by extracting the concepts and explicit relations between them [50]. Examples of a semantic data model include taxonomies, ontologies, and knowledge graphs (KG).

3.1 Ontologies

A colloquial description of an ontology is that of a method to describe concepts and their relationships. Gruber [33] then goes on to define an ontology as a formal, machine-readable, explicit specification of a shared conceptualization characterized by high semantic expressiveness required for increased complexity. Ontologies include abstract concepts or classes, represented as nodes, and predicates representing the relations of these classes (edges in an ontology), with the meaning of the predicates being represented by using rules. Then, individuals

[2] Annex III, European Council position
[3] https://www.openriskmanual.org/ns/doam/index-en.html

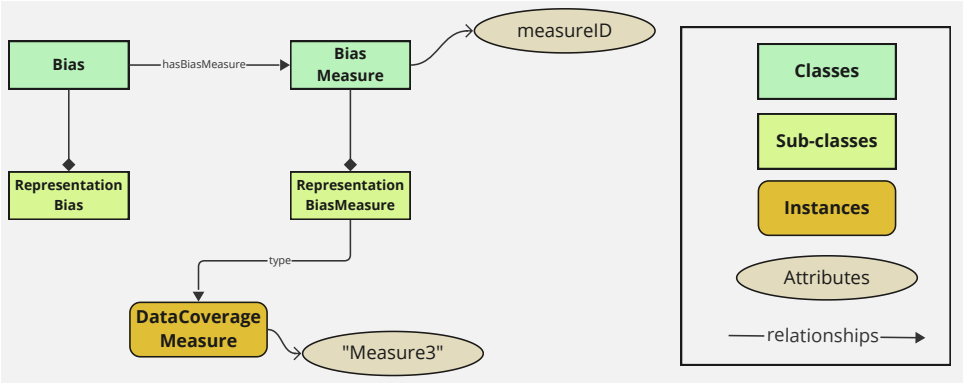


Fig. 2. **Ontology Terminology.** Sample modeling to depict ontology terminology.

or instances are basic components of an ontology. An ontology, together with a set of individual instances, is commonly called a knowledge base. In Figure 2, we illustrate a toy example with "Bias" and "Bias Measure" as **classes**, "has bias measure" as a **relationship** between classes, and "Data Coverage Measure" as an **instance** of the class "Bias Measure".

The use of ontologies (more expressive) or controlled vocabularies (less expressive) allows information on a particular domain to become more aligned and easier to use and share among its expected users [70]. While there is no correct way to model an ontology, they should be designed following objective criteria made up of the following principles: clarity, coherence, extensibility, minimal encoding bias, and minimal ontological commitment [33]. The usefulness of an ontology will depend on the level of agreement on what that ontology defines, how detailed it is, and how widely and consistently it is adopted by the targeted community [36].

3.2 Knowledge Representation Models and Query Languages

Ontologies are specified using knowledge representation models, making the expressiveness of the ontology dependent on the expressive power of the representation model. We list the following models in terms of their expressive power, in increasing order:

- The Resource Description Framework (RDF)[4] enables the description of entities in terms of classes and properties.
- The RDF Schema (RDFs)[5], is an extension of the former, which enables the description of subsumption relations (declaration of hierarchies) of classes between classes, i.e., subclass (see Fig. 2), and properties, i.e., subproperty.
- The Ontology Web Language (OWL)[6] is an ontology language that is formally defined based on description logic. OWL enables reasoning over knowledge-based systems [70], and also makes available a larger number of operators that enable the representation not only of classes, properties, and subsumption relations, but also of class and property constraints, general equivalence relations, and restrictions of cardinality.

[4] https://www.w3.org/RDF/
[5] https://www.w3.org/TR/rdf12-schema/
[6] https://www.w3.org/OWL/

In order to retrieve and manipulate graph data, query languages are used. In the case of our work, we employ the SPARQL query language to analyze data stored in the Resource Description Framework format and perform knowledge discovery.[7]

3.3 RDF Knowledge Graphs

Knowledge graphs (KGs) are data structures that represent factual statements as entities and their relationships using a graph data model [81]. Metadata is part of the KG, as well as taxonomies of entities, relationships, and classes. Ontologies and controlled vocabularies are utilized to describe the meaning of the relations, as well as to annotate entities in a uniform way in the knowledge graph. Thus, a knowledge graph contributes to the development of a common understanding of the meaning of entities in a domain and provides a formal specification of the meaning of these entities.

Several examples of the usefulness of context-aware ontologies for bias awareness and mitigation in ML systems are explored in the work presented in [65]. Further, knowledge graphs, defined using existing ontologies, have gained attention as data structures that allow for the representation of the convergence of data and knowledge in a specific or general domain [34].

The description and modeling of machine learning pipelines and measured biases with ontologies have the capacity to improve the interpretation of bias in data, as well as the understanding of its provenance and context. Furthermore, it will enhance the analysis of the potential detrimental impact that bias in data can have on the overall performance of these pipelines, which can itself support the identification of harms associated with the deployment of AI systems.

4 Scoping through Requirements Gathering and Intended Use

In this section, we will elicit the requirements of our bias ontology by determining the scope in terms of domain identification and the intended use and users of the ontology, employing a general knowledge graph life cycle representation [45, 29].

4.1 Domain Identification: Trustworthy AI and Bias

Given the prevalence of AI in everyday scenarios, recent years have seen the emergence of AI sub fields dedicated to researching multiple ways to build AI systems that incorporate desirable characteristics into their development life cycles [83]. Concepts such as transparency, responsibility, accountability, and fairness are the qualities most often mentioned to describe them in the relevant literature [43]. In addition, it is through the publication of hundreds of academic works and guidelines that seek the development of 'ethical AI' [40] that the conceptualization of the accountability, responsibility, and transparency (ART) framework became prominent, leading to the consolidation of the *Trustworthy AI* framework, pushed largely by regulatory bodies with the aim of guiding commercial AI development to proactively account for ethical, legal, and technical dimensions [32, 55, 76].

Drawing on the emergence of the *Trustworthy AI* framework, alongside it has been a call to establish standards across the field in order to ensure that AI systems are systematically fair and free of bias upon deployment [32]. In this context, the distinction between fairness and bias needs to be made explicit, given that, while often intertwined, they are not always used in conjunction or to study the same phenomena [56, 72, 56].

4.1.1 On Fairness. Fairness in relation to AI systems (i.e., *fair-AI*) emerged as a research area some 15 years ago under the name of discrimination discovery and has evolved to take the form of algorithmic interventions that incorporate mathematical formalizations of moral or legal notions for the fair treatment of different populations

[7] https://www.w3.org/TR/sparql11-overview/

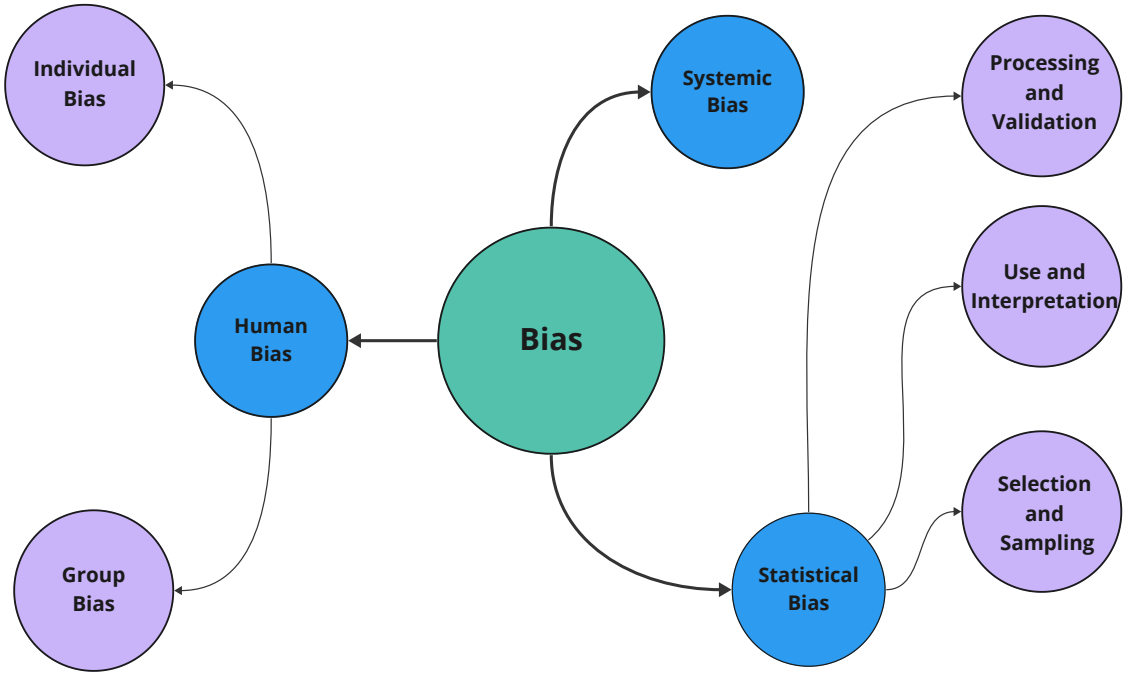


Fig. 3. **Types of Bias and AI Systems.** Core categories of bias defined by NIST.

into machine learning pipelines. These interventions aim to prompt these models to satisfy the statistical non-discrimination criterion for a given subpopulation [11, 6].

4.1.2 On Bias. Bias can be perceived as an "overloaded" concept [56], and has roots in different disciplines, i.e., social science, cognitive psychology, and law [54]. In the context of data analytics and data-driven systems, bias has historically been studied under the lens of the various statistical methods used across all stages of these analyses, i.e., data collection to hypothesis testing. In addressing sociotechnical systems, Friedman & Nissenbaum identified three types of biases that affect them: pre-existing, technical, and emergent biases [27]. In their work, they went on to point out how these biases can produce systemic and discriminatory outcomes.

More recent research on bias and machine learning also demonstrates that bias can start at any point in the ML pipeline and broadens our understanding of how susceptible the entirety of the ML pipeline is to human input. Inspired by the work of Suresh & Guttag, we illustrate an oversimplification of a generic ML life cycle according to our problem formulation (see Fig. 1). In there, we denote how different types of bias can enter this life cycle at any point [79]. For instance, data collection practices involve a series of decisions, such as deciding who is the sampled population, what variables to measure, and the labeling criteria for annotations [79]. The same occurs during model definition and training. For example, a common practice is to use a random seed to preserve reproducibility. Given the stochastic nature of many ML algorithms, the choice of random seed is model-dependent and can significantly alter the outcomes, potentially becoming a source of bias [62].

Consequently, many efforts and resources have been concentrated on devising methods that can improve their identification, understanding, measurement, and mitigation [6].

Notwithstanding, the study of statistical or data biases has come to be conflated with other types of less evident biases, adding to the complexity of analyzing, defining, and identifying all possible biases that datasets and ML models may suffer from [54]. In the same way humans are plagued by innumerable types of bias, datasets and models are also prone to this problem. Moreover, there are competing definitions for the same type of bias, and existing measurements still lack universal consensus [54]. From this perspective, the initiative spearheaded by the National Institute of Standards and Technology (NIST) introduces one of the first formal guidances and standards that support the identification and management of bias in AI pipelines [71]. Following an extensive literature review, interviews with experts, and public consultation, the resulting report proposes a thorough, but not exhaustive, categorization of different types of bias identified in relation to AI pipelines that are beyond common computational definitions. In Figure 3, we illustrate the three main categories of bias according to the NIST classification: statistical, systemic, and human bias, serving as a starting point for a crescent repository.

The explicit identification and definition of different types of biases are helpful in expanding our common understanding of these phenomena and their interplay with other components of an AI pipeline and stakeholders and can also contribute to improving the identification of harms stemming from the application of AI systems in everyday scenarios.

4.2 Identifying Stakeholders and Sources of Domain Terminology

Following the identification and description of the domain we will be focusing on with our modeling, we move on to identify key stakeholders and to gather more relevant sources of terminology. Doing so allows us to specify a use case characterized primarily by the actors involved in the intended usage of the ontology, as per their tasks, needs, and roles.

4.2.1 Stakeholders. In order to identify stakeholders, we take advantage of our own position and involvement in a research project on bias in relation to AI systems from 2020-2023.[8] The research agenda of this project set out to address the whole AI decision-making pipeline with the overall goal of understanding the different sources of bias, detecting them as they manifest, and mitigating their effects on the produced results for specific applications. In particular, the work presented here contributes to deepening the understanding of the impact of bias in data; further, by employing ontological formalisms and semantic data models, we set out to enable the production of documentation artifacts to further the efforts of the research community to instill accountable practices in the context of AI development.

Resorting to the context of this project and its collaborative nature, it is possible for us to hold regular and fruitful discussions with experts researching different dimensions of bias from a multidisciplinary and critical point of view; see [66, 6]. These exchanges with researchers also help deepen our understanding and characterization of bias in data from a critical stance (e.g., there is never just one bias, bias detection is contextual, bias detection can depend on data modality, and biases cannot be eradicated) and identify challenges not only in modeling bias but also in relation to the underlying documentation process, primarily on how it should not be fully automatized.

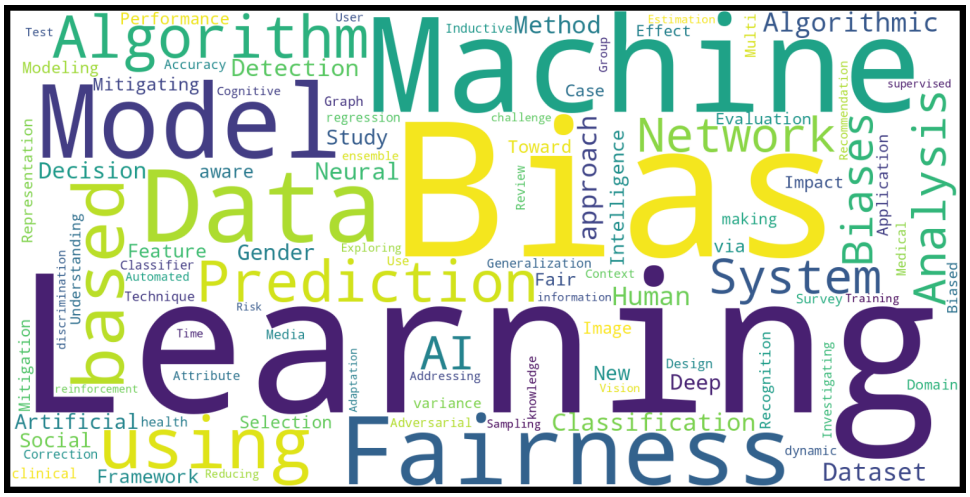
In developing a bias vocabulary in the context of Trustworthy AI, it is important to aim for a careful balance between an effective, useful, and comprehensive vocabulary that supports streamlining documentation tasks while, at the same time, avoiding dissuading practitioners from critical thinking when engaging in both documentation and bias analysis. The aim of both practices is to mitigate the negative consequences arising from the deployment of ML systems. However, it is always possible that, unintentionally through the enforcement of standardization or automation on practitioners, new gateways are created that worsen the problem. Some influencing factors include lack of experience, domain knowledge, or the existence of the right incentives [10, 49, 35].

[8] https://nobias-project.eu/

```
( TITLE ( survey ) AND TITLE ( bias ) AND TITLE-ABS-KEY ( data-driven OR "artificial intelligence" OR
"machine learning" ) ) AND PUBYEAR > 2019 AND PUBYEAR < 2025
```

Show less

(a) Sample Search Query



(b) WordCloud Top 50 Most Frequent Terms in Abstracts

Fig. 4. **Specifications from Search of Academic Articles in Scopus.**

Ultimately, this rapport informs our design choices across all iterations of ontology engineering, makes us aware of the limitations of our technical tool, and creates opportunities for refinement in later versions.

4.2.2 Sources of Terminology. The development of our vocabulary is informed by the growing body of literature on our domain of interest in order to complement the NIST formal guidance. In this case, we particularly rely on periodically identifying relevant literature in order to gather more background information and to employ as sources for terms, definitions, and relationships. The aim is also to employ these resources to excerpt relevant terminology that might not be modeled in relevant domain ontologies considered for reuse. Ultimately, the objective is to produce a rich vocabulary of biases in relation to Trustworthy AI.

During the first iteration of our search, we first focus on manually curating a list of emerging work on bias published at venues such as ACM FAccT,[9] AAAI/ACM AIES,[10] ACM EAAMO.[11] However, in efforts to standardize our search, we select academic articles by querying literature databases for survey papers published during

[9] https://facctconference.org/
[10] https://www.aies-conference.com/
[11] https://eaamo.org/

Table 1. **Relevant Concepts in the Trustworthy AI Domain and their Relationships to Bias.**

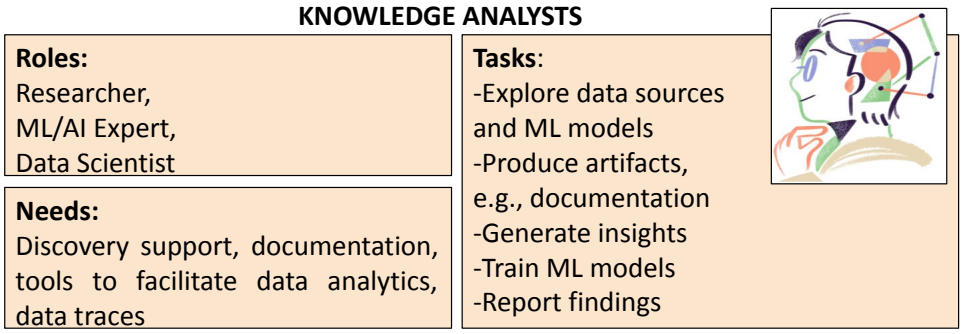
Concept	Definition	Relationship
Bias	A concentration on, or interest in, one particular area or subject. Whilst a more value-laden definition conceptualizes bias as prejudice for or against one person or group, especially in a way considered to be unfair [56].	-
ML Problem	An ML problem or task is the formal description of a process that needs to be completed (e.g., based on inputs and outputs [58].	is evaluated for
Dataset	A collection of data, published or curated by a single source, and available for access or download in one or more representations [58].	is evaluated for
Bias Measure	A quantitative metric or indicator that assesses the presence and extent of bias in a particular context via predefined thresholds [23].	measures
Application	The use, purpose, or application of an ML-powered system. Examples include recommenders, speech recognition, etc.	is associated to
Harm	Adverse lived experiences resulting from the deployment of AI systems and their operation in the world [10, 73].	aligns

2019-2024; given the large body of literature available, we opt for prioritizing survey papers of relative recency. We pay attention to the discerning of bias and its detection measures from fairness notions and their measures, combining keywords such as "survey," with "bias" and "machine learning" or "artificial intelligence."

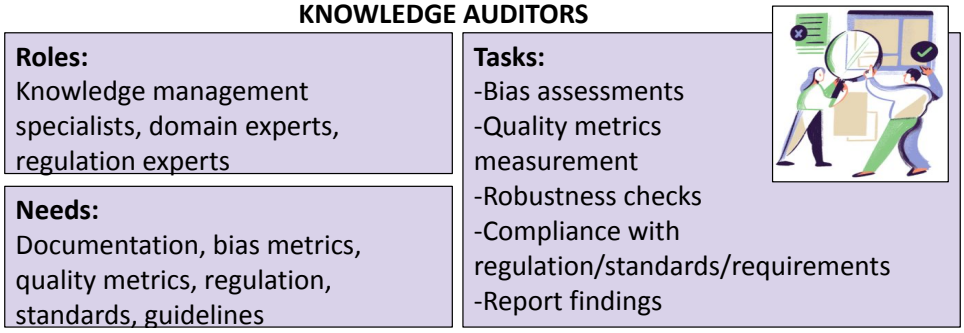
We run our first set of queries over SCOPUS (33 results),[12] and then we run the same set of queries over the ACM Digital Library (24 results).[13] We do so in order to contrast the results obtained, with manual checks asserting similarities for the results obtained; at the same time, we review for relevancy, filtering out irrelevant papers. In future iterations of our work, we look forward to adopting a data-driven approach for terminology extraction as a more systematized strategy for vocabulary generation, given the limitations of our manual approach. In Figure 4a, we include a sample search query in the database, and in Figure 4b, we illustrate the generated WordCloud from the top 50 most frequent terms present in abstracts extracted from relevant documents.

Prevalent terms and concepts in the Trustworthy AI domain that are relevant for our modeling include bias, machine learning, dataset, task, application, fairness, harms, and risks. Through our scoping process, we also learn more about how these concepts interact or relate to each other. In Table 1, we summarize the principal concepts that serve as a starting point to further develop our vocabulary. In the table, we also include the relationship of these concepts with regard to bias. Each concept we identify represents the topmost abstract concept in a hierarchy of terms, with less abstract or more concrete concepts defined as the bias vocabulary grows to give a

[12] https://www.scopus.com/home.uri
[13] https://dl.acm.org/



(a) Knowledge Analyst



(b) Knowledge Auditor

Fig. 5. **Knowledge Graph Ecosystem Actors.** Two distinct KGE actors: KG Analyst (Figure 5a), KG Auditor (Figure 5b). The intended users of the bias ontology play different roles with specific needs to accomplish the required tasks of that role.

broader coverage. For example, `Bias` is the most abstract representation, while `Representation Bias` is a more concrete type of bias.

4.2.3 Intended Use and Users. The last step in scoping the bias ontology is defining its intended use and intended users to facilitate modeling. As already alluded to, the primary intended use of our ontology is to semantically represent knowledge on bias detection assessments. This includes the representation of types of bias and associated bias measures as they are detected in relation to the components of the machine learning pipeline. We also wish to represent additional knowledge that relates to the individual components, such as descriptions of datasets, as well as other aspects of bias detection, such as sociotechnical harms. The objective is to improve the contextual analysis and understanding of bias for AI researchers and practitioners. We employ the representation of the KG management life cycle to delve deeper into the ontology's main users or actors, their roles, and their needs.

Actors, Roles, and Needs. Ontologies are components of knowledge graph ecosystems (KGE); other components also include data sources and knowledge graphs [29]. These components are subjected to a knowledge management life cycle, comprised of different steps that enable their "creation, validation, curation, maintenance, traversal, and analysis" [29]. In relation to these steps, actors with their corresponding roles, tasks, and needs emerge. [29] defines these concepts as such:

- *Actors* are individuals responsible for contributing to the execution of the life cycle steps. The five actors defined in [29], include knowledge providers, knowledge builders, knowledge auditors, knowledge analysts, and knowledge consumers.
- *Roles* are played by an actor based on their relation with the knowledge graph ecosystem.
- *Needs* are the combination of requirements and constraints stated by an actor on the basis of the role they are playing.
- *Tasks* are the functions to be performed by an actor under a particular role.

From these definitions, we elaborate further on the following actors relevant to the identified users of the bias ontology: knowledge analysts and knowledge auditors.

Knowledge Analysts have a direct interaction with the components of the KGE, their goal being to gain insights from it; analysts can include data scientists or ML/AI experts, both in an academic setting or industrial setting, looking to enhance their understanding of bias, contextualized to a particular ML/AI problem they are working on. **Knowledge Auditors** have a different type of interaction with components of the KGE, as they assess them based on different dimensions, such as quality or compliance with predefined requirements and needs. Auditors are usually individuals with expertise in the domain, technical specifications, and the relevant regulation. In this context, we refer to experts aiming to provide insights into the functioning of an AI system. In Figure 5, we schematize these two actors based on their roles, needs, and tasks in relation to the bias detection assessments use case.

5 Modeling and Implementation

In this section, we describe the design stages of Doc-BiasO. We also describe its implementation and include examples of instances.

5.1 Modeling Doc-BiasO

To model our ontology, we adhere to best practices for ontology engineering [33, 42]: definition of competency questions, identification of reusable ontologies, and establishment of annotation conventions.

5.1.1 Competency Questions Definition. The competency questions that emerged from the analysis phase showcase the intended use of Doc-BiasO, first as part of documentation frameworks, by providing the vocabulary needed to describe AI pipelines. And then, to provide AI researchers or practitioners with a resource that informs them on how bias interplays with other components in data or when they are researching the development of a new measure and wish to survey existing ones. We enumerate the set of defined questions in Table 2.

5.1.2 Reusable Ontologies Identification. Reusable ontologies are identified following a layered approach [42]:

(1) a foundational layer for general metadata and provenance;
(2) a domain-dependent layer to cover standards for the relevant area of use, e.g., machine learning systems;
(3) a domain-dependent layer of ontologies specific to our problem of interest, e.g., bias in data measurements.

We **first** lay the foundation of our ontology by reusing ontologies such as:

- the SKOS data model [50], which allows us to express basic structures for concept schemes;
- the PROV data model (PROV-O) [44], as it enables the representation of provenance information generated by different entities and can be easily applied to multiple contexts;
- the Friend of a Friend (FOAF) vocabulary [84], in order to use its collection of terms to describe claims about different things, i.e., people, groups, and documents.

The **second** layer incorporates standard schemas for data mining and machine learning algorithms, such as the Machine Learning Schema (MLS) ontology [58]. This schema provides fine-grained descriptions to represent the characteristics and intricacies of ML models. Similarly, the Data Catalog Vocabulary (DCAT) [2] enables the

Table 2. **Competency Questions.**

Nº	Competency Question
Q1	Given a particular bias, what is its definition?
Q2	Given a particular bias, what are the AI applications that could be associated with it?
Q3	Given a particular bias, what AI harm could be aligned with it?
Q4.1	Given a particular bias, how many measures have been documented for it?
Q4.2	Given a particular bias, what measures have been documented for it?
Q5	Given a bias measure, in which scholarly document was it defined?
Q6	Given a bias measure, what is its formalization?
Q7	Given a bias measure, what dataset feature is evaluated by it?
Q8	Given a bias measure, what machine learning task is being evaluated by it?
Q9	Following its implementation, what is the score of the evaluated bias measure?
Q10	What is the amount of documentation generated across the ML pipeline?

fine-grained description of datasets and data services in a catalog using a controlled and rich vocabulary. By extension, the Data Quality Vocabulary (DQV) [4] provides a framework and vocabulary to assess the quality of a dataset, offering an extensive catalog of quality metrics.

For the **third** layer, we look at previous work on bias, specifically the BODP [41] and the Artificial Intelligence Ontology (AIO).[14] The class `AIO:Bias` is our starting point, which we organize resorting to hierarchies via `rdfs:SubClassOf` both as per the AIO modeling and also to represent different kinds of bias identified in the literature, i.e., representation bias, popularity bias, and demographic bias. We build on the design pattern and AIO ontology, but since it does not satisfy our modeling needs, missing concepts are manually incorporated, as we set out to capture and explicitly document otherwise unstated assumptions about bias in relation to ML systems [17].

Critical data studies maintain that for bias detection tasks to be meaningful, practitioners must *reflect* on possible harms that can arise after deploying an ML system in dynamic social and cultural contexts [73, 17]. Here, we emphasize on the importance of assisting practitioners via the development of tools that streamline tasks that may be perceived as a burden [49], while avoiding dissuading them from reflecting on the harms that could emerge from the implementation of these systems. For that reason, in our modeling we align scoped biases with harms, with the objective of making explicit the articulation of otherwise alleged, unstated negative consequences attributed to ML systems. However, our expectation and recommendation is that users will enrich the proposed vocabulary with the results derived from their own explorations, in a similar line as with AI incident databases.[15]

Furthermore, bias is not singular and is highly context dependent, meaning that most biases are studied and defined in association with a particular ML application. To represent both of these concepts, we model `bias:Harm` and `bias:Application`. The central concept in our ontology is `bias:BiasMeasure`. This class represents a measure defined in some `foaf:Document`, evaluated in a `dcat:Dataset` (that has some characteristics), and for a particular `mls:MLTask`. `bias:BiasEvaluation` is the class that represents the n-ary relationship between entities schematized in the extended entity relationship model completed at the beginning of the design phase.

[14] https://bioportal.bioontology.org/ontologies/AIO?p=summary
[15] https://oecd.ai/en/catalogue/tools/ai-incident-database

5.1.3 Annotation Convention Establishment. We operate under the *minimal completeness principle* concerning existing metadata [42]. All components represented in Doc-BiasO have information such as a label, comment, or definition. We aim to include a source for any given definition.

5.2 Towards a Comprehensive Vocabulary for Trustworthy AI

The Trustworthy AI framework requires a comprehensive formal vocabulary that unifies approaches and contemplates terminology and concepts of ML pipelines and, in broader terms, AI holistically. This type of resource can contribute to the generation of metadata that primes the reproducibility and traceability of research results [53, 46], a known issue in ML research and development [59, 3]. Moreover, it can help achieve a certain degree of standardization in the area.

Motivated by this, we perform a thorough analysis of the FMO ontology [25, 26], and the VAIR vocabulary [31], itself an extension of the AIRO ontology [30], to determine their characteristics and how they fit into our modeling as we identify domain-dependent ontologies. We also do this with the aim of achieving a good balance between ontology reuse and down-the-line overhead derived from doing so [42].

The results of this analysis are:

(1) FMO complements Doc-BiasO by providing coverage of existing fairness metrics used to evaluate ML systems. Specifically, metrics related to machine learning problems of classification and regression. At the time of writing, there is also an extension to give coverage to clustering problems under development;
(2) VAIR captures a wider scope of AI system deployment to instill accountability on an AI provider (i.e., a party that places the system on the market) and thus capture specifications of risky applications of AI from a regulatory point of view. Specifically, it expresses risk as per the EU AI Act and key standards in the ISO 31000 series;
(3) both ontologies represent overlapping concepts, e.g., algorithm, dataset, and ML systems. Particularly, FMO reuses vocabularies such as MLOnto[16], and the OBO ontology;[17] and
(4) both of these ontologies represent bias, however, with differing modeling objectives. FMO organizes `fmo#Bias` in a hierarchy with eight subclasses. The class `fmo#Fairness Notion` "measures" `fmo#Bias`. VAIR represents `vair#Bias` as a subclass of `airo#Consequence`. Figs.6-7, correspondingly, illustrate the graphs generated by the OntoGraph plugin for Protégé for the `Bias` class in each ontology.

Moreover, we perform an alignment analysis for both ontologies. We start by using an implementation of the LogMap ontology matching service [39], which relies on lexical and structural indexes to enhance scalability and complete any needed semantic disambiguation by manual curation. An overview of the matching results obtained from alignment analysis is included in Table 3 and the resulting integrated and curated ontology can be accessed on GitHub.[18]

After this analysis, we conclude that for the development of our ontology, it is not favorable to import either ontology in its entirety. We do this in order to avoid compromising or constraining our modeling, in particular as we notice here the potential for duplicate concepts with overlapping or related definitions, a problem that is more prominent when developing a vocabulary following a bottom-up approach [42]. In our case, FMO and VAIR, both model `Bias`, as does the AIO ontology. Here, we did not create a new class but extended AIO, and when needed, we implement OWL axioms to assert class equivalence, i.e., `owl:equivalentClass`. Otherwise, we reference external concepts using annotation properties or explanatory notes that expand class definitions.

[16] https://osf.io/chu5q/
[17] http://obofoundry.org
[18] https://github.com/SDM-TIB/Doc-BIASO

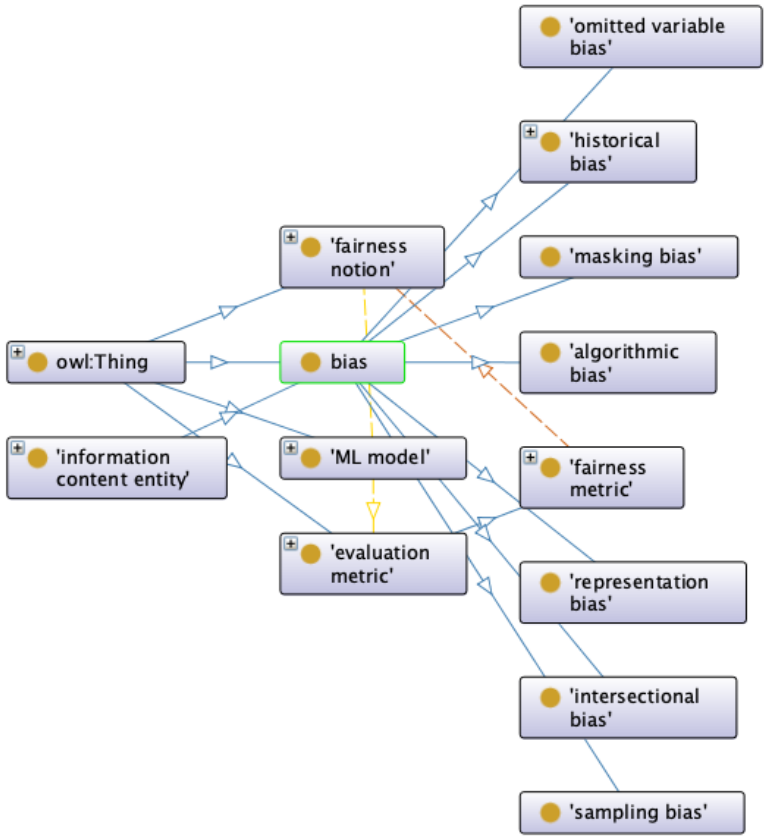


Fig. 6. **Bias Modeling in the FMO Ontology [25, 26].**

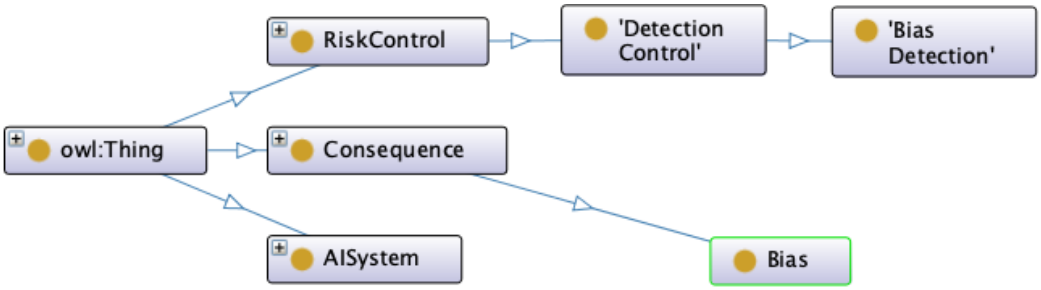


Fig. 7. **Bias Modeling in the VAIR Vocabulary [31].**

Table 3. **Alignment Study Results.** Example mappings between FMO and VAIR.

Entity 1	Entity 2	Relation	Result
fmo: machine learning model	vair: Machine Learning Model	≡	Correct
fmo: bias	vair: Bias	≡	Correct
mlo: Algorithms	airo: Algorithm	≡	Correct
obo: STATO _0000415	vair: Low Accuracy	≡	Incorrect
fmo: separation class fairness notion	vair: Sound Source Separation	≡	Incorrect
fmo: ML dataset	vair: Dataset	≡	Manually created

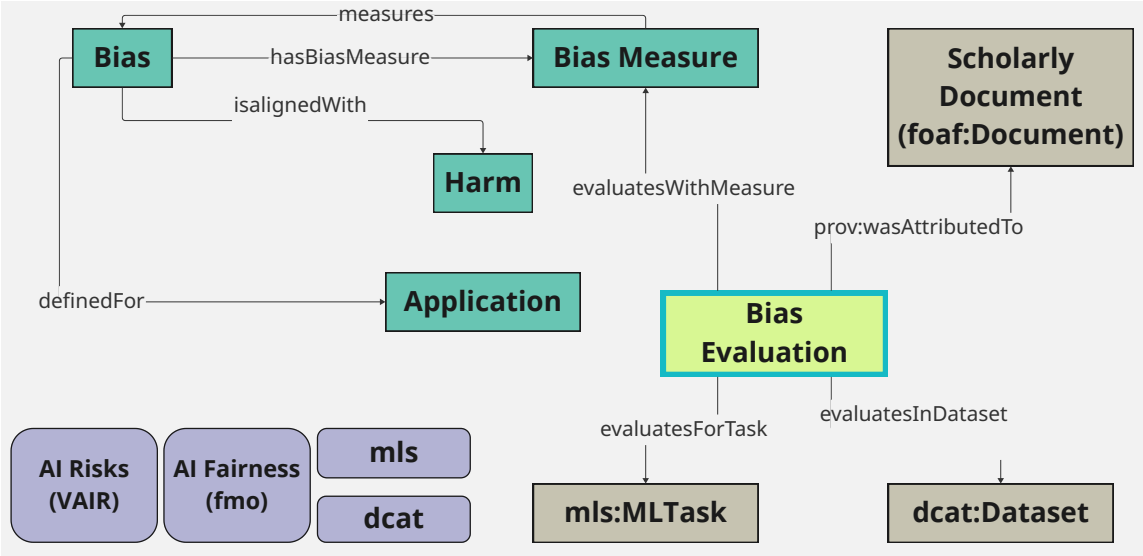


Fig. 8. **Conceptualization of the Doc-BiasO Ontology.** Core concepts in the ontology are represented as classes, in color-coded boxes, to account for originating vocabularies. Concisely, in teal are new classes, and in khaki we denote reused classes and state the originating prefix. Object properties are drawn as directed arrows between classes. The names of relevant, prominent, and reused ontologies and taxonomies are highlighted in purple-colored boxes.

5.3 Doc-BiasO Specifications

Figure 8 illustrates a conceptual overview of the principal classes and relationships of Doc-BiasO. We use rectangular boxes to illustrate classes. In teal, we denote classes created for Doc-BiasO. The class `Bias Evaluation` is highlighted in brighter green, as this class represents an n-ary relationship in our original schematization. The khaki rectangles represent classes from reused vocabularies, such as Friend of a Friend (FOAF), ML Schema (MLS), and the Data Catalog Vocabulary (DCAT). Finally, we use purple rectangles with rounded edges to prominently highlight reused ontologies and taxonomies.

Table 4 summarizes the number of reused concepts.

Table 4. **Reused Concepts.** Summarization of the number of reused concepts from existing vocabularies and ontologies.

Originating Vocabulary Suffix	Reused Concepts
FOAF+	70
AIO	64
mls	27
DQV	13
SKOS	6
DCAT	4
FOAF	3
PROV-O	2
DOAM	2
BIAS-ODP	1

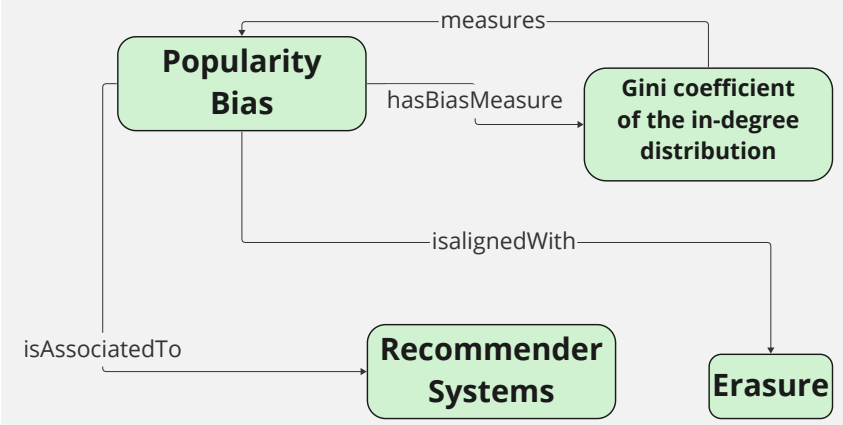


Fig. 9. **Conceptualization of an Instance of Doc-BiasO.** Instances of the Doc-BiasO ontology are represented with round-edge boxes and the color green. "Popularity Bias" is an instance of bias:Bias. Related classes are also exemplified.

5.3.1 Doc-BiasO Axiomatization. The conceptualization of the Doc-BiasO ontology is specified using OWL logical axioms. In doing so, we enable consistency checks and logical inferences on a resulting RDF knowledge graph. The exemplification of some domain range axioms for the Bias and Bias Evaluation classes, as well as axioms denoting restrictions on Bias expressed in OWL can be found in Appendix A.

5.3.2 Instantiating Doc-BiasO. To showcase an instantiation of Doc-BiasO, we take a look at an example based on bias detection in relation to recommender systems, commonly implemented in online social networks.

The class `Bias` is instantiated as `Popularity Bias`. This bias `is Associated With` an instance of the class `Application`, `Recommender System` and `has a Bias Measure`, "`Gini coefficient of the in-degree distribution`". In this same example, `Popularity Bias is Aligned With` the instance of the class `Harm`, which is `Erasure`. Figure 9 illustrates this example.

5.3.3 Doc-BiasO in Numbers and Access. The current version of the Doc-BiasO ontology is made up of 390 classes, 72 object properties, and 28 data properties. It is publicly available as a VoCol repository[19] supported by TIB.[20] VoCol provides an interface for querying the ontology and also enables the visualization and exploration of the ontology. The metadata describing each of its components can also be accessed at VoCol.

6 Evaluation

This section includes the results of the technical evaluation of our ontology. Specifically, we (1) measure the coverage of represented knowledge; (2) implement competency questions expressed in natural language as SPARQL queries; and (3) carry out an automatic assessment using various online tools.

6.1 Ontology Coverage Evaluation

Given that there are no official specifications or benchmarks established that can provide us with a good measure for domain coverage in terms of bias, we perform a comparative analysis between Doc-BiasO, FMO, and VAIR. Following this, we adhere to ontology design best practices and identify a preliminary set of four use cases to populate our ontology over the identified core concepts.

Table 5. **Comparison of Ontologies.** Comparative between the Fairness Metrics Ontology (FMO), Vocabulary of AI Risks (VAIR) and Doc-BiasO, in the context of bias coverage.

Ontology Name	Nº Classes	Bias Coverage	Metrics
Fairness Metrics Ontology (FMO)	1239	Bias is related to a fairness notion that is measured in relation to a dataset, given a fairness metric defined in the literature.	Fairness Metrics, Statistical Metrics to Evaluate Datasets
Vocabulary of AI Risks (VAIR)	424	Bias is a subclass of consequence in relation to AI risks	Out of scope
Doc-BiasO	390	Bias in the central class in the ontology, supporting the measurement of different types of it with specific metrics defined in the literature.	Bias Metrics, Dataset Characteristic Taxonomy (MLSO-DC), Evaluation Measure Taxonomy (MLSO-EM)

6.1.1 Ontology Comparison: FMO, VAIR, and Doc-BiasO. In order to facilitate a better understanding of the differences, similarities, and strengths of our proposal, in Table 5, we include a comparative view of our ontology in relation to two relevant ontologies, the Fairness Metrics Ontology (FMO) and the Vocabulary of AI Risks (VAIR). On the table, it is possible to see the size of all three ontologies and the considerations of each with regard to how they model bias and metrics. As we have already alluded to, bias in FMO is related to a fairness notion that is measured in relation to a dataset, given a fairness metric. In the context of VAIR, bias is a subclass of consequence in relation to AI risks; in this case, metrics are out of scope in any case. While we acknowledge

[19] http://ontology.tib.eu/DocBIASO/visualization
[20] TIB – Leibniz Information Centre for Science and Technology and University Library

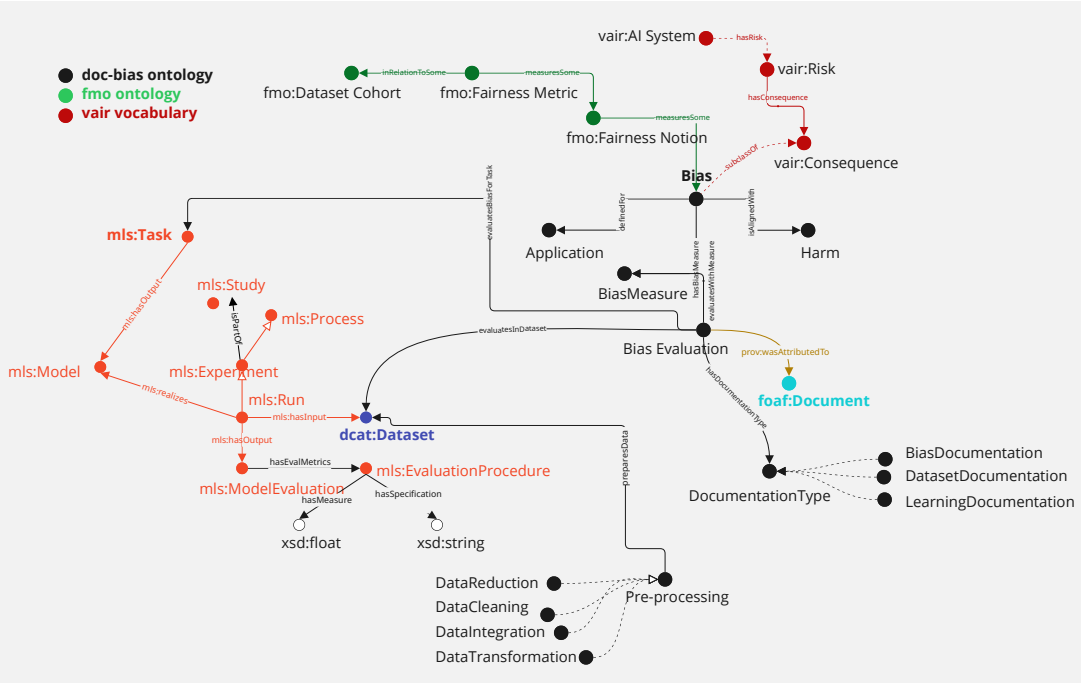


Fig. 10. **Domain Coverage Visualization.** Doc-BiasO, FMO, and VAIR are modeled together to display their complementary nature and a suggested integration.

the seeming similarities between measuring fairness and bias, however, we emphasize that by focusing on bias, it is possible to widen the scope of analysis across the machine learning pipeline. In order to provide a better approximation and appreciation as to the coverage of Doc-BiasO, and to better illustrate how all three ontologies are complementary and primed for an integrated use, we include Figure 10.

6.1.2 Representing Knowledge on Bias to Improve Coverage. Our first use case is the NIST Special Publication "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence" [71]. This publication includes the categorization of 51 types of bias. In our ontology, we model all 51 subclasses of bias adhering to the suggested hierarchy; additional types of bias that emerge from the literature not already included in the NIST categorization are added (i.e., Intrinsic Bias and Extrinsic Bias). Figure 11 illustrates the Bias hierarchy modeled in Doc-BiasO.

For the remaining three use cases chosen to populate our ontology, we resort to existing literature as per our scoping strategy (see Section 4.2). We select three survey papers from the obtained documents; we specifically choose these papers as they were part of the results obtained from querying both databases. Additionally, the papers survey literature on bias metrics for different ML problems, covering different types of bias. In doing so, our ontology will contain varied empirical examples for bias metrics modeling.

The chosen survey papers were:

(1) Survey on Bias and Fairness in Machine Learning [47]. The authors of this paper surveyed more than 50 papers targeting bias, with the objective to taxonomize and summarize the current (at the time of publishing) state of research in algorithmic biases in relation to machine learning across different areas, e.g., classification, regression, and clustering, specifically looking at bias in data and algorithms;

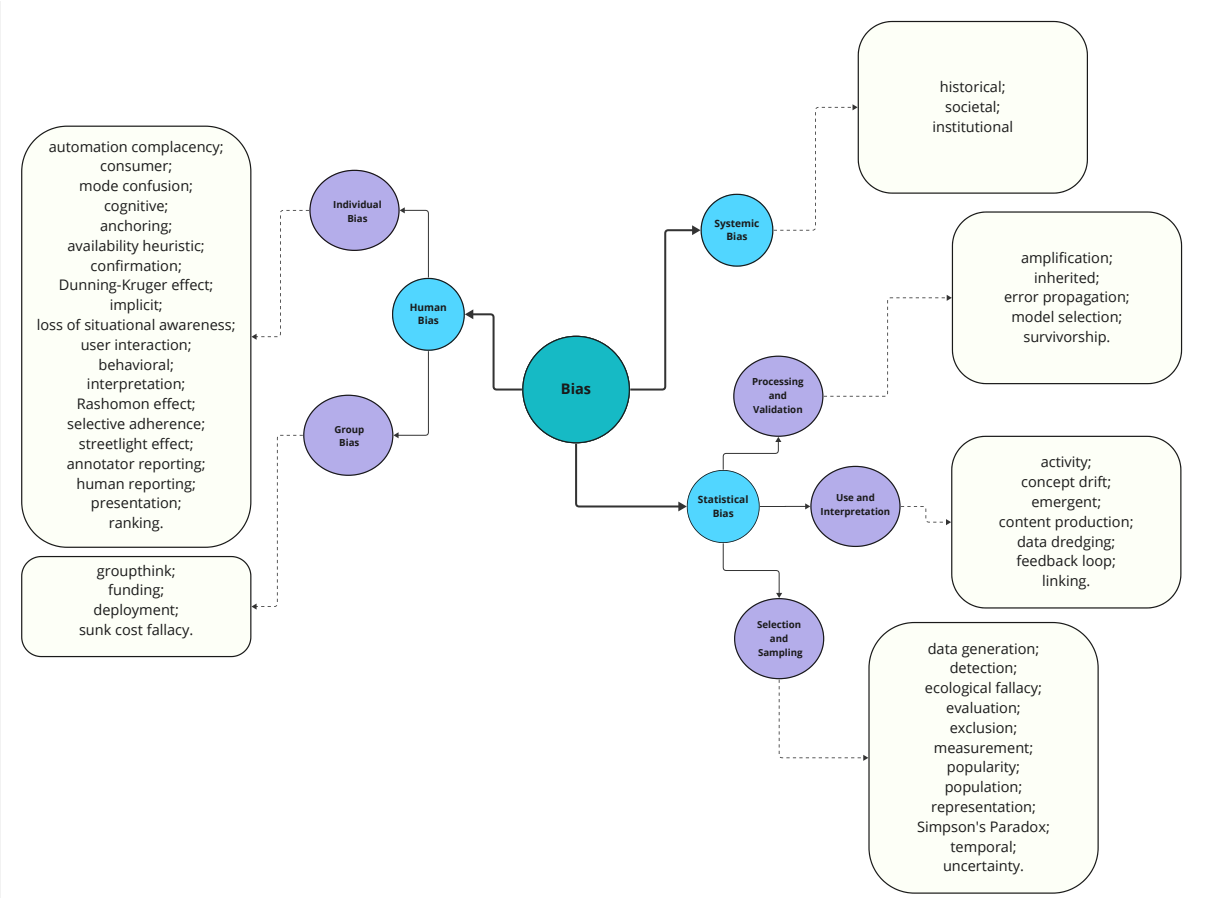


Fig. 11. **Bias Hierarchy as per NIST.** The 51 types of biases categorized in the NIST publication and modeled in Doc-BiasO.

(2) Representation Bias in Data: A Survey on Identification and Resolution Techniques [72]. The authors of this paper also surveyed more than 50 papers with a focus on bias in data, specifically literature identifying representation bias as a feature of a dataset across structured and unstructured data. We use this paper as a case study given the prevalence of representation bias as a feature of all datasets, independent of the machine learning problem; and

(3) Measuring Fairness with Biased Rulers: A Comparative Study on Bias Metrics for Pre-trained Language Models [20]. The authors of this paper surveyed more than 30 papers that compare bias metrics for contextualized language models. We use this paper as a case study given the rise of popularity in bias detection in natural language processing resources, an area that continues to present challenges for bias detection and mitigation efforts given the current popularity of large language models.

In Figure 12, we show a capture of the instances modeled in Doc-BiasO corresponding with bias measures extracted from the mentioned survey papers.

As part of the evaluation process, we also report on the completeness, or domain coverage, of Doc-BiasO in terms of bias and bias measures and summarize our results in Table 6.

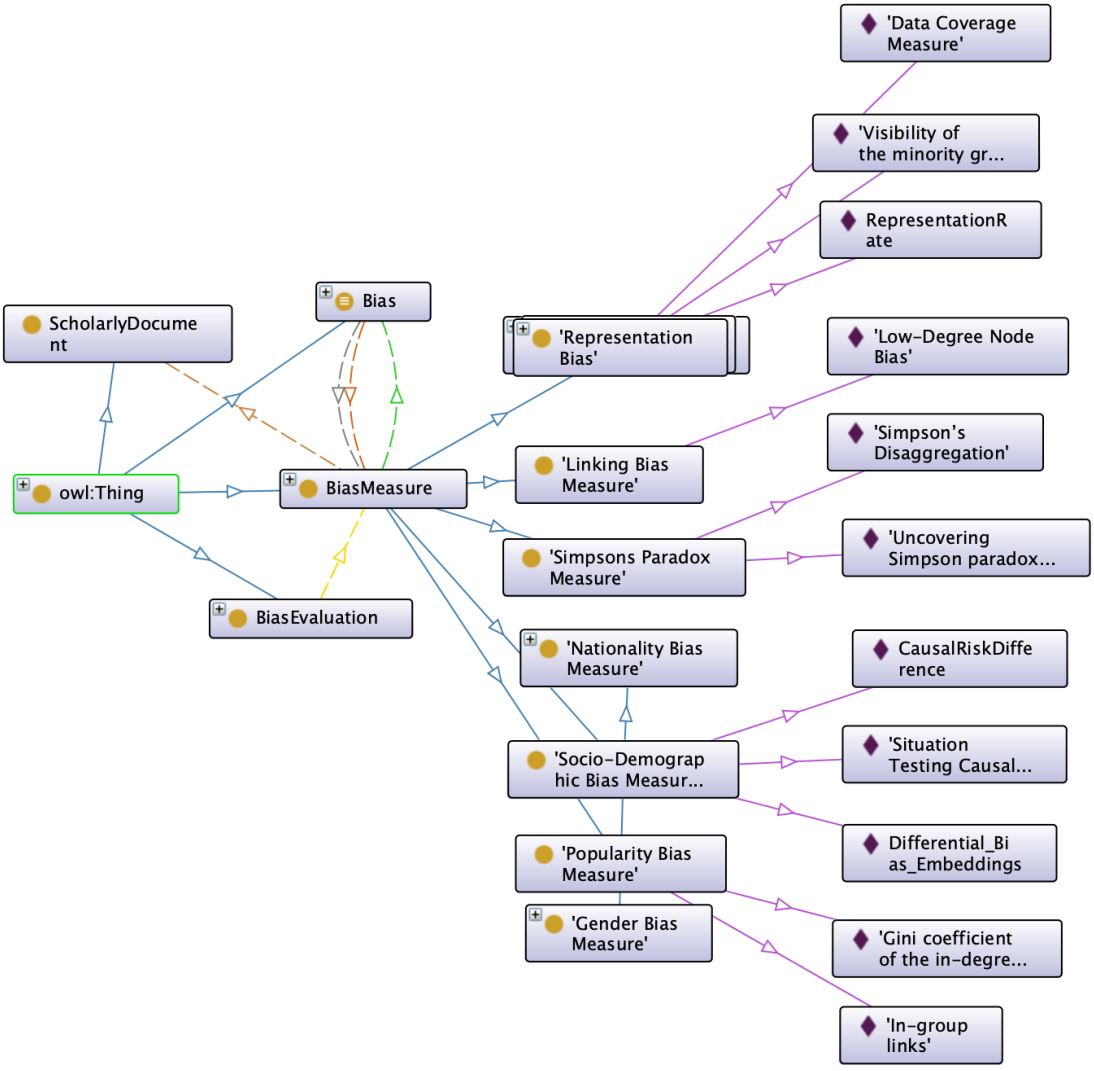


Fig. 12. **Doc-BiasO in Granular Detail.** Modeled instances of the Bias Measure class.

Table 6. **Represented Knowledge on Bias of Doc-BiasO.**

Indicator	Results
Completeness	
Bias	All 51 subclasses have verifiable definitions based on the NIST report, $\frac{59}{51}$ = 115%.
Bias Measures	8 subclasses with verifiable definitions based on ongoing literature review, 24 instances based on 3 case studies.

6.2 Competency Question Evaluation

The domain analysis and scope definition of Doc-BiasO, as already described in Section 5, derived a set of competency questions that was also used to convey the requirements that would guide the engineering of our ontology. As part of the process, we tested and refined the Doc-BiasO ontology by implementing the formalization of the competency questions originally expressed in natural language as SPARQL queries.[21] The queries were tested to make sure the results were the expected ones.

```
PREFIX  skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX  owl:  <http://www.w3.org/2002/07/owl#>
PREFIX  rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX  bias: <https://bias-project.x/bias/>

SELECT DISTINCT
    ?bias_1 (COUNT(DISTINCT ?biasMeasure_1) AS
    ?number_of_measures)
WHERE { ?bias_1  rdfs:subClassOf bias:Bias .
        ?biasMeasure_1 bias:measures ?bias_1}
GROUP BY ?bias_1
```

Listing 1. SPARQL Query for Competence Question Q4.1

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX bias: <https://bias-project.x/bias/>

SELECT DISTINCT

    ?biasMeasure_1 ?definition_1  ?formalization_1
WHERE {
   ?biasMeasure_1 rdfs:subClassOf bias:BiasMeasure ;
       skos:definition ?definition_1 ;
      bias:formalization ?formal_1
FILTER ( (  REGEX(str(?biasMeasure_1), "Gini", 'i'))))}
```

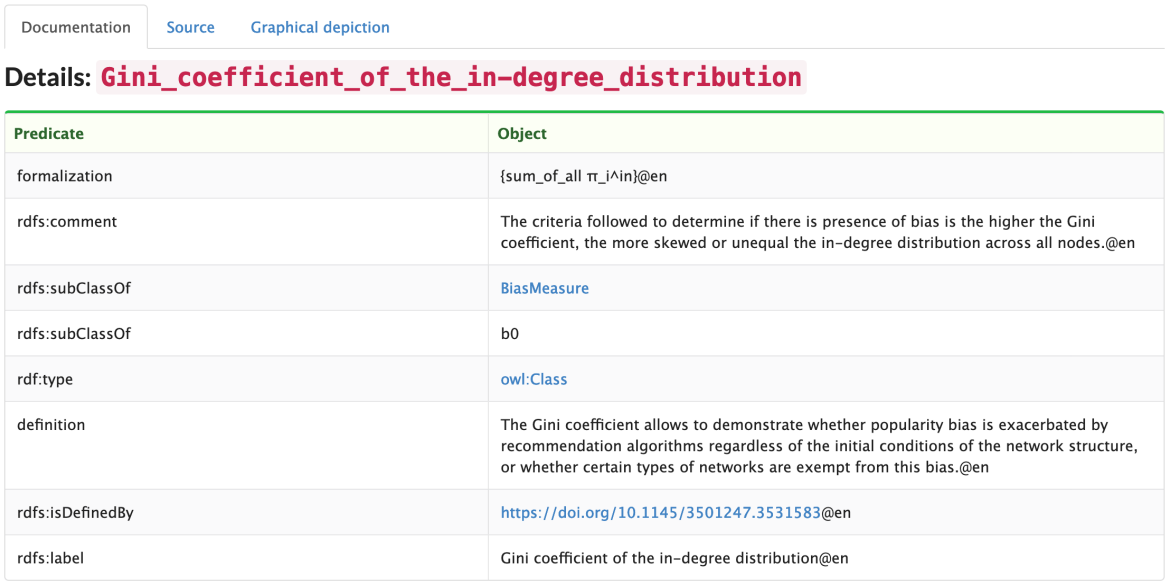
Listing 2. SPARQL Query for Competence Question Q6

To illustrate their adequacy, we continue with the example introduced in Section 5, and start by posing Q1 *"Given a particular bias, what is its definition?"*; our example uses `Popularity Bias`. Below is the query result:

> "When collaborative filtering recommenders emphasize popular items (those with more ratings) over other "long-tail," less popular ones that may only be popular among small groups of users."@en

This expected result is expressed as a `rdfs:Literal` in `English`. We follow this question by posing Q4.1 *"How many measures have been documented for it?"*, as specified by the corresponding query in Listing 1. The execution of this query yields the bias type and the number of measures. In this case and at the time of evaluation, `Popularity Bias` has 2 measures. We then choose the measure `Gini coefficient of the in-degree distribution` to learn more about it. We proceed to execute the query that corresponds to Q6. *"what is its formalization?"*. The corresponding SPARQL query is specified in Listing 2, and additional metadata produced for it is illustrated in Figure 13, containing the definition for the chosen measure and the formalization for it in natural language.

[21] https://www.w3.org/TR/sparql11-query/



Documentation | Source | Graphical depiction

Details: `Gini_coefficient_of_the_in-degree_distribution`

Predicate	Object
formalization	{sum_of_all π_i^in}@en
rdfs:comment	The criteria followed to determine if there is presence of bias is the higher the Gini coefficient, the more skewed or unequal the in-degree distribution across all nodes.@en
rdfs:subClassOf	BiasMeasure
rdfs:subClassOf	b0
rdf:type	owl:Class
definition	The Gini coefficient allows to demonstrate whether popularity bias is exacerbated by recommendation algorithms regardless of the initial conditions of the network structure, or whether certain types of networks are exempt from this bias.@en
rdfs:isDefinedBy	https://doi.org/10.1145/3501247.3531583@en
rdfs:label	Gini coefficient of the in-degree distribution@en

Fig. 13. **Bias Metric Metadata Example.** Metadata for Gini coefficient of the in-degree distribution (CQ6).

6.3 Automatic Ontology Evaluation

This version of Doc-BiasO has been validated with online tools to verify its consistency and syntactical validity and check for modeling anomalies or errors.

First, we checked that our ontology is syntactically correct using the W3C RDF validation service.[22] The results indicated a successful validation of our RDF document. Secondly, we checked for logical consistency by running the DL reasoning engine Pellet (v.2.2.0), as a plug-in for the Protégé open-source platform (v.5.6.1).[23] We choose this engine, as it is a complete reasoner. The results determined that Doc-BiasO is logically coherent and consistent. Finally, we scanned our ontology with the "OOPS! Ontology Pitfall Scanner" [57] to automatically dismiss the existence of modeling pitfalls; the evaluation results were also positive, as there were no bad practices detected by the tool.

7 Doc-BiasO in Use

In this section, we show how to use Doc-BiasO to document biases in data. First, we describe a use case that resorts to implementing representation bias measures over a benchmark dataset, through the perspective of the *Knowledge Analyst*. Second, we describe a use case that employs Doc-BiasO as part of the implementation of a neuro-symbolic system to document bias across the ML pipeline [68]; here we emulate the perspective of the *Knowledge Auditor*, given the granularity of detail the documentation manages to achieve.

7.1 Use Case: Age and Representation

Here, we describe the first use case for Doc-BiasO, which resorts to the implementation of representation bias measures over a benchmark dataset.

[22] https://www.w3.org/RDF/Validator/
[23] https://protege.stanford.edu/software.php

7.1.1 Benchmark Dataset. The datasets we use here were elaborated from data of the United States (US) Census, and were introduced as an updated version of the UCI Adult dataset [13]. The American Community Survey (ACS) Public Use Microdata Sample (PUMS) - ACS PUMS - Dataset [22], comprises tabular data on individuals across the United States spanning multiple years; the addition of a temporal and geographical dimension facilitates richer analysis and benchmarking across seemingly similar subpopulations. Here, we use data pertaining to the states of California and Florida for 2018. The datasets are accessible through the Folktables Python package.[24]

7.1.2 Bias Measures. We perform our analysis on the instance `Representation Bias` of the class `Bias`. The variable studied is age, and the aim is to perform a bias analysis, assessing if individuals are underrepresented or not according to their age group membership. A detailed study as such can be useful, as age discrimination in the context of employment is rampant, and the design of inclusive private or public policies for hiring, training, and other employment conditions should account for a heterogeneous and dynamic population with divergent needs. In particular, one of the emergent harms aligned with this bias is `Erasure`:

> "Erasure or social invisibility refers to a group of people in society that can be excluded or systematically ignored from resource allocation procedures or decision-making processes due to data collection practices or historical record keeping."@en

For the analysis, we implement two measures for the same bias and annotate the described dataset based on the values obtained. This enables a comparison of the results for the differing measures. Additionally, it allows us to underpin the importance of a human in the loop when it comes to choosing the appropriate measure and determining thresholds, as well as the interpretation of the results.

We define the following measures as instances of the class `Bias Measure`:

- **Data Coverage**, or having enough similar entries for each object in a dataset [72]. Given a dataset D, with an attribute of interest x, a count threshold τ (e.g., $\tau = 100$), and subgroups g (e.g., age=middle-age, age=young-adult) defined over x. The dataset satisfies coverage over g if there are at least τ objects in D that correspond to the subgroup g of x (e.g., there are more than 100 instance counts for both the age = middle age and age = young adult subgroups).
- **Representation Rate**, or having an equal number or a representative percentage of objects for different subgroups in a dataset in relation to a majority subgroup [72]. Given a dataset D of d dimensions, following a distribution of p: $\Omega \rightarrow [0,1]$, with an attribute of interest x, and a threshold $\tau \in (0,1]$, (e.g., $\tau = 0.67$); D is said to have a representation rate τ with respect to x if for all subgroups i, j of g, it yields $\frac{n_i}{n_j} \geq \tau$. τ close to 0 indicates bias in D.

7.1.3 The Doc-BiasO KG. The RDF knowledge graph that presents our use case is generated over a data integration system, DIS=$\langle O, S, M \rangle$ [38], where O corresponds to the Doc-BiasO ontology, containing concepts and properties as described in previous sections. S is the set of data sources that populate the graph, e.g., benchmark datasets and bias measures execution, and M comprises mapping rules expressing correspondences between O and S specified in the RDF Mapping Language (RML) [21]. The resulting RDF Knowledge Graph (accessible via a public GraphDB instance) is defined by 35 RML rules and has 4,819 statements.

7.1.4 Drawing Insights. We extract the count of individuals for each of the age groups defined according to the US Census Bureau[25], and also perform a distribution analysis over the datasets for California (CA) and Florida (FL). This overview already let us identify the age group that is more prominent in our dataset, Age group 4, comprising ages in the range 45-64. Although age group 6, which encompasses ages in the range 85-90, is the

[24] https://github.com/socialfoundations/folktables
[25] https://www.census.gov/

Table 7. **Value Counts for Age Variable** On the left, the results for California, and on the right, Florida.

Age Range	Age Group	CA	FL
16 - 24	Age group 1	49 215	21 212
25 - 34	Age group 2	51 602	22 510
35 - 44	Age group 3	47 141	22 130
45 - 64	Age group 4	101 029	58 315
65 - 84	Age group 5	54 396	42 393
85 - 90	Age group 6	4 603	3 425

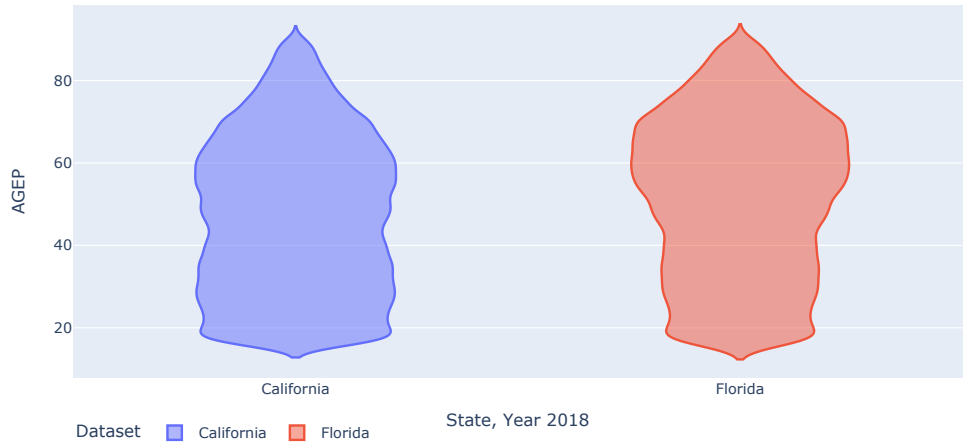


Fig. 14. **Distribution of Age Variable in CA & FL.**

smallest. In addition, the age between states denotes a similar distribution. These results are summarized in Table 7 and Figure 14.

```
PREFIX  bias: <https://bias-project.x/bias/>
PREFIX  rdf: <http://www.w3.org/rdf-syntax-ns#>
SELECT DISTINCT ?bias_eval_1 ?bias_measure ?data
?feature ?score
WHERE {
    ?bias_eval_1 rdf:type bias:BiasEvaluation ;
    bias:evaluatesWith ?bias_measure;
    bias:evaluatesInDataset ?data ;
    bias:evaluatesFeature ?feature ;
    bias:biasClassification ?score }
```

Listing 3. SPARQL Query for Competence Question Q9

We then extract the results of the bias evaluation (see the SPARQL query shown in Listing 3). The analysis of representation rate is performed by making pairwise comparisons between subgroups, while data coverage is

less constrained, assessing for a minimum count of instances established at $\tau = 100$ in relation to the majority subgroup. For this reason, data coverage yields that all age groups are adequately represented in both datasets in relation to Age Group 4. Whilst the representation rate yields that both datasets are always biased against Age group 6, the pairwise comparisons denote that the CA dataset is also always biased against Age group 4, whilst the FL dataset is always biased against Age group 1. The distribution of the results is reported in Figure 15.

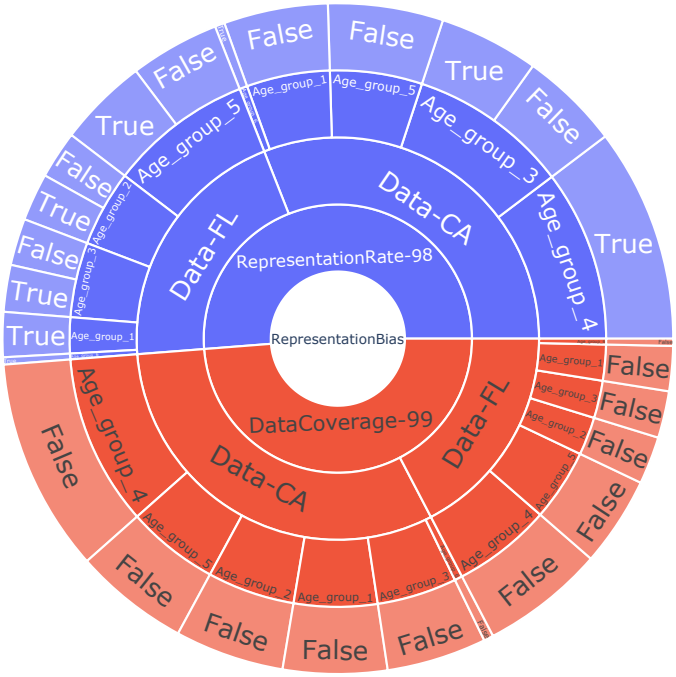


Fig. 15. **Analysis of Results.** Data distribution for two measures for Representation Bias for different age groups in CA & FL.

7.2 Use Case: Neuro-symbolic System to Document Bias in ML Pipelines

The integration of sub-symbolic and symbolic systems into neuro-symbolic systems is seen as one possible way to remedy the "black-box" problem associated with many modern ML-powered systems [80, 68]. In order to characterize hybrid learning systems, boxology design patterns, as introduced in [80], provide building blocks for the combination of symbolic and sub-symbolic architectures. As an example, Figure 16 illustrates a design pattern for explainable learning systems through rational reconstruction. The proposed design denotes the combination of an ML model that is first trained to then infer predictions and then passes through a reasoning system (e.g., the classification outcomes). The paired pattern depicts how the model and predictions are semantically enriched with background knowledge (see 'model:semantic' on Fig. 16). This background knowledge includes the definition of a data integration system and bias measures; the execution of the data integration system results in a knowledge

graph (KG) that comprises the results of tracing the ML pipeline and measuring bias (i.e., symbol: trace). Queries can be executed over the KG to recover the data required for bias analysis and to uncover patterns that may explain the effect of bias in data on the decisions made by the sub-symbolic system, e.g., an ML system.

As a practical example, we show how we integrate Doc-BiasO in the implementation of a neuro-symbolic system instantiated for the problem of misinformation classification. Albeit simple, this use case puts into perspective the versatility of our ontology and the benefits of producing traces of the ML pipeline as factual statements in a KG that are human- and machine-readable. This is also something that positions our tool for ML pipeline inspection at a fine-grained level to facilitate the tasks of the *Knowledge Auditor*.

The documentation system and use case we describe in this section are a natural extension of the work presented in this paper, which has been adapted from [68].

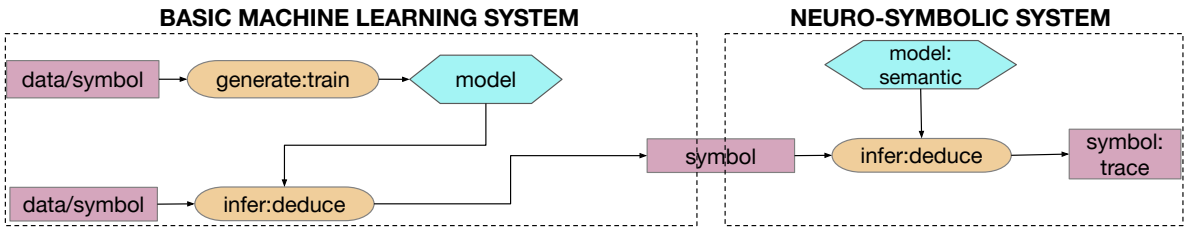


Fig. 16. **Boxology Design Patterns [80].** Design pattern for a neuro-symbolic architecture to improve interpretability of ML systems. A statistical model is integrated with a symbolic model and background knowledge to support the interpretation of the output generated.

7.2.1 Implementation and Overview of Resources. As already alluded to, the implementation of the neuro-symbolic system also relies on the declarative definition of a data integration system, $\langle O, S, M \rangle$. This facilitates the transformation and integration of the different data sources to create an RDF knowledge graph, as well as facilitates tracking the semantic enrichment obtained across the pipeline. In order to do so, we define three Documentation Steps: data ingestion, learning and output, and bias assessment. The integration system we use here is composed of 1) an instantiation of Doc-BiasO, O; 2) data sources pertaining to the entire ML pipeline, thus including datasets and data on the training and inference process. S; and 3) sets of mapping rules that align the data in the source with the concepts in the ontology, M.

Benchmarks: The datasets used in the implementation of our framework are part of the FakeNewsNet catalogue[26] published by Shu et al. [74]. They are the BuzzFeed dataset and the PolitiFact dataset.

Model: Probabilistic Soft Logic (PSL) is a statistical programming language that falls under the realm of statistical relational learning frameworks [7], which are known for their effectiveness at defining probabilistic models over complex relational data, combining graphical models and first-order logic [75]. Rules are weighted with scores that represent the importance of each rule; they are learned during the training phase. Specifically, Chowdhury et al. [19] resort to PSL to specify the rules that should guide a fake news classifier based on their joint-credibility score model (CSM).

Bias Measures: With the objective to provide a comprehensive framework that enables the trace and analysis of bias in data-driven systems, we implemented the following set of measures over the knowledge graph: overrepresentation metric, similarity metric, and frequency measure.

[26] https://github.com/KaiDMML/FakeNewsNet/tree/old-version

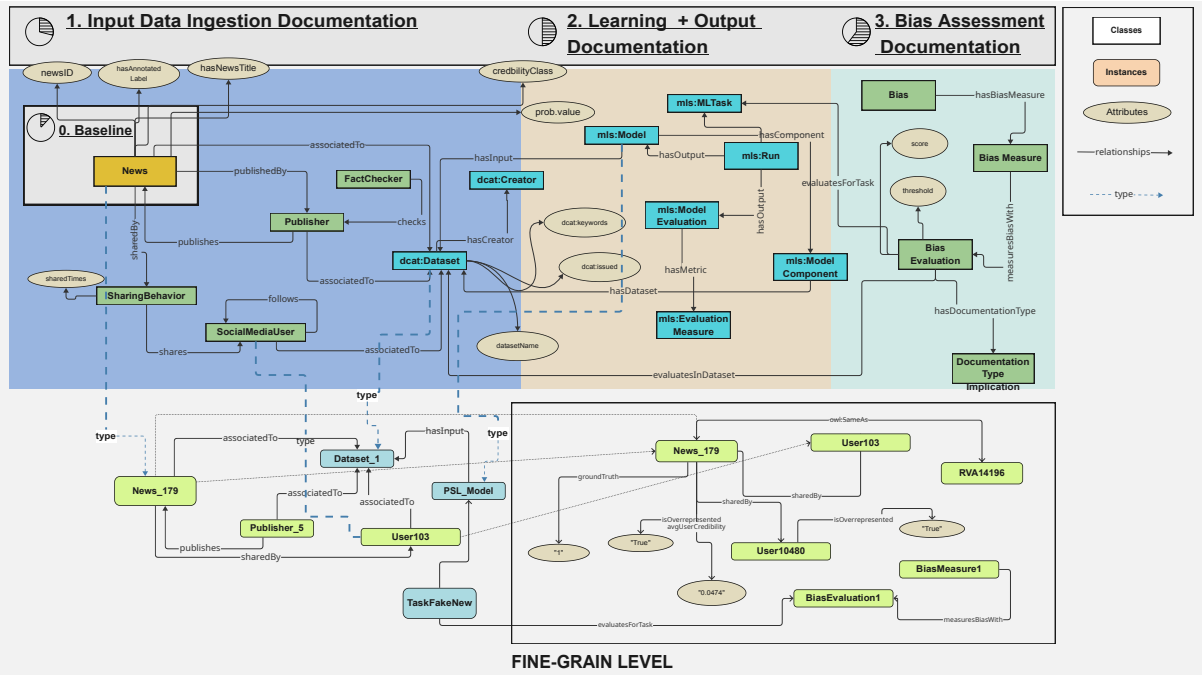
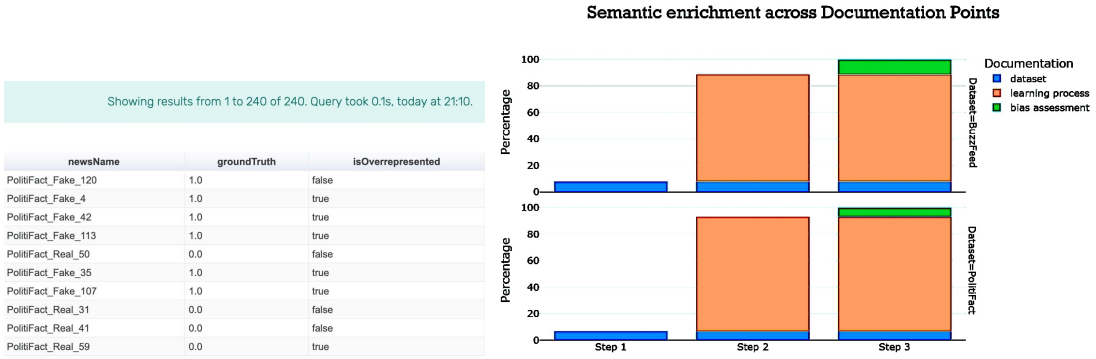


Fig. 17. **Documenting a Machine Learning Pipeline.** Above, the metadata schema of Doc-BiasO is depicted. Classes, characteristics, and relationships among them are modeled to provide the necessary background knowledge to trace the ML pipeline. Documentation Steps are also illustrated across the pipeline, alluding to the gain in semantic enrichment derived from describing the pipeline, measured in comparison to 0.Baseline. There, the target entity of class 'News' is identified. Doc-BiasO prompts the elaboration of machine-readable and FAIR documentation artifacts at coarse- and fine-grain levels.

7.2.2 Documenting Bias with Doc-BiasO and a Neuro-Symbolic System. Figure 17 provides a conceptualization of Doc-BiasO integrated in a neuro-symbolic system. In the upper part of the illustration, we have an overview of an instantiation of Doc-BiasO for the context of misinformation detection that defines classes, i.e., news, publishers, (social media) users, datasets, and ML models, as well as their attributes, and the relationships between classes, i.e., shares (news), follows (user), and publishes (news). Below this, we illustrate how the data is enriched by these semantic representations, thus encoding information in a machine-readable format on the results of implementing the overrepresentation metric on entities generated during the learning step. This prompts the generated KG to produce bias-aware documentation artifacts at coarse- and fine-grain levels (see Figure 18a for a snippet example of generated output).

The real-world usability of the output generated by our neuro-symbolic documentation system relies on the inherent role these systems play in enhancing the interpretability of sub-symbolic systems. The fine-grain output on bias generated by the neuro-symbolic system can thus be used to gain new insights based on bias patterns in the data previously unobservable, employing, for instance, data visualization techniques that can overcome mere abstractions [45, 16]. In our example, entity overrepresentation derives from a data imbalance created in the preprocessing step; our output can inform on the need to adopt a strategy that can alleviate this problem depending on the access to the original model, e.g., by data resampling or model recalibration; if there is no access to the model, the output can be used to elaborate on the limitations of the model and its results.

Our documentation system allows us to measure the degree of semantic enrichment gained across the Documentation Steps of the classification pipeline. Figure 18b illustrates the number of instances (in relative terms) in accordance with each of the steps for both datasets. The amount of metadata generated for the "learning and output" step is significantly larger in comparison to the other components; this is because we managed to generate and describe data for the whole sub-symbolic system and integrate it into our documentation pipeline.



Showing results from 1 to 240 of 240. Query took 0.1s, today at 21:10.

newsName	groundTruth	isOverrepresented
PolitiFact_Fake_120	1.0	false
PolitiFact_Fake_4	1.0	true
PolitiFact_Fake_42	1.0	true
PolitiFact_Fake_113	1.0	true
PolitiFact_Real_50	0.0	false
PolitiFact_Fake_35	1.0	true
PolitiFact_Fake_107	1.0	true
PolitiFact_Real_31	0.0	false
PolitiFact_Real_41	0.0	false
PolitiFact_Real_59	0.0	true

(a) Output Snippet (b) Semantic Enrichment Results

Fig. 18. **Results of Documentation System.** Left, a snippet of machine- and human-readable documentation generated for results of applying the overrepresentation bias metric on entities during the learning process. Right, the percentage of semantic enrichment across each documentation step: input data ingestion, learning and output, and bias assessment [68].

We ultimately propose a documentation approach that sets out to support KG Auditors and KG Analysts to further elucidate the impact of bias pipelines beyond the analytical capabilities of existing frameworks for bias analysis despite additional work being needed to fully meet the different expectations of users' when it comes to a suitable user interface.

Moreover, from a technical and practical side, there are additional open challenges. First, sub-symbolic systems are inherently constrained to the characteristics of the data they are trained on [59]. Additionally, technical and domain knowledge are needed to capture and document the intricacies of a particular pipeline, as compared to documenting at a coarse-grain level. With regard to the neuro-symbolic system generation, another factor to consider is how obtaining a fine-grained level of detail in terms of generated metadata will also incur costs in terms of compute and data storage; thus, for each documentation task, finding a balance between efficiency and effectiveness is essential.

8 Conclusions and Future Work

In this work, we presented Doc-BiasO, an ontology for bias measures found in the literature that can support the elaboration of documentation of bias in machine learning pipelines. Our objective is to contribute towards improving the interpretation of these pipelines in terms of biases captured and the derived harms attributed to ML systems. Further, we make a call for a unified controlled vocabulary for the Trustworthy AI framework and assess existing relevant work. We technically evaluated Doc-BiasO and presented two examples as to how to use it from the perspective of the *Knowledge Analyst* and the *Knowledge Auditor*. The results show how Doc-BiasO can be used to document representation bias in regard to age in a popular benchmark through the implementation of two measures. We also report on the intricacies of doing so, and while we only use two states for our use case, the analysis can be easily extended to integrate more datasets, as well as the predictive model and produced

output. The results of the second example show a comprehensive way of documenting bias across the whole ML pipeline, employing a neuro-symbolic system. This denotes how our tool supports the creation of fine-grained documentation by measuring semantic enrichment of target entities in a machine learning problem.

Notwithstanding, our work is not without limitations. Firstly, research on bias in machine learning, and by extension AI, is a fast-moving field; thus, providing adequate and updated coverage with our tool is a challenge. Secondly, bias evaluations are highly complex and context-dependent tasks. This means that our modeling cannot account for all potential existing biases and that, in general, bias analysis cannot be fully automated, requiring a human-in-the-loop. Thirdly, our resources are yet to be evaluated by ML practitioners outside a research environment. Nevertheless, the addressed limitations are an opportunity for future work. In particular, we intend to add and expand on aspects left unmodeled in this version with regard to bias measures, and we will liaise with ML practitioners to evaluate the suitability of our tool in real-world scenarios. We will also continue the development of a controlled vocabulary for Trustworthy AI, as this resource can foster effective communication between the different actors involved across the ML pipeline.

Acknowledgments

Mayra Russo wishes to thank Guillermo Climent-Gargallo, Sammy Sawischa, and Yukti Sharma for their support during this investigation. Mayra Russo received support by the EU-Horizon 2020 research and innovation programme under the MCSA grant agreement No. 860630, project: NoBIAS. Maria-Esther Vidal is partially supported by the Leibniz Association under the "Leibniz Best Minds: Programme for Women Professors", TrustKG-Transforming Data in Trustable Insights; Grant P99/2020. The views reflected on this work are those of the author only, and the European Research Executive Agency is not responsible for any use that may be made of the information it contains.

References

[1] F. AI. 2021. Fairness flow. https://ai.facebook.com/blog/how-were-using-fairness-flow-to-help-build-ai-that-works-better-for-everyone/. (2021).

[2] R. Albertoni, D. Browning, S. J. D. Cox, et al. 2023. The w3c data catalog vocabulary, version 2: rationale, design principles, and uptake. *ArXiv*, abs/2303.08883.

[3] R. Albertoni, S. Colantonio, P. Skrzypczynski, et al. 2023. Reproducibility of machine learning: terminology, recommendations and open issues. *ArXiv*, abs/2302.12691.

[4] R. Albertoni and A. Isaac. 2020. Introducing the data quality vocabulary (dqv). *Semantic Web*, 12, 81–97.

[5] A. Aler Tubella, D. Coelho Mollo, A. Dahlgren Lindstrom, et al. 2023. Acrocpolis: a descriptive framework for making sense of fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '23). ACM, Chicago, IL, USA, 1014–1025. ISBN: 9798400701924. DOI: 10.1145/3593013.3594059.

[6] J. M. Alvarez, A. B. Colmenarejo, A. Elobaid, et al. 2024. Policy advice and best practices on bias and fairness in AI. *Ethics Inf. Technol.*, 26, 2, 31. DOI: 10.1007/S10676-024-09746-W.

[7] S. H. Bach, M. Broecheler, B. Huang, et al. 2017. Hinge-loss markov random fields and probabilistic soft logic. *Journal of Machine Learning Research*, 18, 109, 1–67. http://jmlr.org/papers/v18/15-631.html.

[8] R. Baeza-Yates. 2020. Bias on the web and beyond: an accessibility point of view. In *W4A '20: 17th Web for All Conference, Taipei, Taiwan, April 20-21, 2020*. C. Duarte, T. Drake, F. Hwang, and C. Lewis, (Eds.) ACM, 10:1. DOI: 10.1145/3371300.3385335.

[9] R. Baeza-Yates. 2020. Biases on social media data: (keynote extended abstract). In *Companion of The 2020 Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*. A. E. F. Seghrouchni, G. Sukthankar, T. Liu, and M. van Steen, (Eds.) ACM / IW3C2, 782–783. DOI: 10.1145/3366424.3383564.

[10] A. Balayn, M. Yurrita, J. Yang, et al. 2023. "fairness toolkits, a checkbox culture?" on the factors that fragment developer practices in handling algorithmic harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '23). ACM, Montréal, QC, Canada, 482–495. ISBN: 9798400702310. DOI: 10.1145/3600211.3604674.

[11] S. Barocas, M. Hardt, and A. Narayanan. 2019. *Fairness and Machine Learning*. http://www.fairmlbook.org. fairmlbook.org.

[12] S. Barocas and A. D. Selbst. 2016. Big data's disparate impact. *California Law Review*, 104, 671.

[13] B. Becker and R. Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20. (1996).

[14] A. Beer, M. Brunet, V. Srivastava, et al. 2022. Leibniz data manager - A research data management system. In *The Semantic Web: ESWC 2022 Satellite Events - Hersonissos, Crete, Greece, May 29 - June 2, 2022, Proceedings* (Lecture Notes in Computer Science). P. Groth et al., (Eds.) Vol. 13384. Springer, 73–77. DOI: 10.1007/978-3-031-11609-4_14.

[15] E. M. Bender and B. Friedman. 2018. Data statements for natural language processing: toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6, 587–604. DOI: 10.1162/tacl_a_00041.

[16] A. Bigelow, K. Williams, and K. E. Isaacs. 2021. Guidelines for pursuing and revealing data abstractions. *IEEE Transactions on Visualization and Computer Graphics*, 27, 2, 1503–1513. DOI: 10.1109/TVCG.2020.3030355.

[17] S. L. Blodgett, S. Barocas, H. Daume III, et al. 2020. Language (technology) is power: a critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, (July 2020), 5454–5476. DOI: 10.18653/v1/2020.acl-main.485.

[18] J. Buolamwini and T. Gebru. 2018. Gender shades: intersectional accuracy disparities in commercial gender classification. In *FAT*.

[19] R. Chowdhury, S. Srinivasan, and L. Getoor. 2020. Joint estimation of user and publisher credibility for fake news detection. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*. M. d'Aquin, S. Dietze, C. Hauff, E. Curry, and P. Cudré-Mauroux, (Eds.) ACM, 1993–1996. DOI: 10.1145/3340531.3412066.

[20] P. Delobelle, E. Tokpo, T. Calders, and B. Berendt. 2022. Measuring fairness with biased rulers: a comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, (Eds.) Association for Computational Linguistics, Seattle, United States, (July 2022), 1693–1706. DOI: 10.18653/v1/2022.naacl-main.122.

[21] A. Dimou, T. D. Nies, R. Verborgh, et al. 2016. Automated metadata generation for linked data generation and publishing workflows. In *Proceedings of the Workshop on Linked Data on the Web, LDOW 2016, co-located with 25th International World Wide Web Conference (WWW 2016)* (CEUR Workshop Proceedings). Vol. 1593. CEUR-WS.org. https://ceur-ws.org/Vol-1593/article-04.pdf.

[22] F. Ding, M. Hardt, J. Miller, et al. 2021. Retiring adult: new datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34.

[23] C. Dwork, M. Hardt, T. Pitassi, et al. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*. ACM, 214–226. DOI: 10.1145/2090236.2090255.

[24] I. Fernandez, C. Aceta, E. Gilabert, et al. 2023. Fides: an ontology-based approach for making machine learning systems accountable. *Journal of Web Semantics*, 79, 100808. DOI: https://doi.org/10.1016/j.websem.2023.100808.

[25] J. S. Franklin, K. Bhanot, M. Ghalwash, et al. 2022. An ontology for fairness metrics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '22). Association for Computing Machinery, Oxford, United Kingdom, 265–275. ISBN: 9781450392471. DOI: 10.1145/3514094.3534137.

[26] J. S. Franklin, H. Powers, J. S. Erickson, et al. 2023. An ontology for reasoning about fairness in regression and machine learning. In *Iberoamerican Conference on Knowledge Graphs and Semantic Web*.

[27] B. Friedman and H. Nissenbaum. 1996. Bias in computer systems. *ACM Trans. Inf. Syst.*, 14, 330–347.

[28] T. Gebru, J. Morgenstern, B. Vecchione, et al. 2021. Datasheets for datasets. *Commun. ACM*, 64, 12, 86–92. DOI: 10.1145/3458723.

[29] S. Geisler, C. Cappiello, I. Celino, et al. 2025. From genesis to maturity: managing knowledge graph ecosystems through life cycles. *Proc. VLDB Endow.*, 18, 3, (Sept. 2025), 1390–1397. DOI: 10.14778/3718057.3718067.

[30] D. Golpayegani, H. J. Pandit, and D. Lewis. 2022. AIRO: an ontology for representing AI risks based on the proposed EU AI act and ISO risk management standards. In *Towards a Knowledge-Aware AI - SEMANTiCS 2022 - Proceedings of the 18th International Conference on Semantic Systems, 13-15 September 2022, Vienna, Austria* (Studies on the Semantic Web). A. Dimou, S. Neumaier, T. Pellegrini, and S. Vahdati, (Eds.) Vol. 55. IOS Press, 51–65. DOI: 10.3233/SSW220008.

[31] D. Golpayegani, H. J. Pandit, and D. Lewis. 2023. To be high-risk, or not to be—semantic specifications and implications of the ai act's high-risk ai applications and harmonised standards. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '23). Association for Computing Machinery, Chicago, IL, USA, 905–915. ISBN: 9798400701924. DOI: 10.1145/3593013.3594050.

[32] H. L. E. Group. 2019. Ethics guidelines for trustworthy ai. en. (2019). DOI: https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html.

[33] T. R. Gruber. 1995. Toward principles for the design of ontologies used for knowledge sharing? en. *International Journal of Human-Computer Studies*, 43, 5-6, (Nov. 1995), 907–928. DOI: 10.1006/ijhc.1995.1081.

[34] C. Gutierrez and J. F. Sequeda. 2021. Knowledge graphs. *Commun. ACM*, 64, 3, 96–104. DOI: 10.1145/3418294.

[35] A. K. Heger, L. B. Marquis, M. Vorvoreanu, et al. 2022. Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata. *Proc. ACM Hum.-Comput. Interact.*, 6, CSCW2, Article 340, (Nov. 2022), 29 pages. DOI: 10.1145/3555760.

[36] A. Hogan, E. Blomqvist, M. Cochez, et al. 2021. Knowledge Graphs. *ACM Comput. Surv.*, 54, 4, (July 2021), 1–37. arXiv: 2003.02320. DOI: 10.1145/3447772.

[37] I. Hupont, D. Fernández-Llorca, S. Baldassarri, et al. 2024. Use case cards: a use case reporting framework inspired by the european ai act. *Ethics and Information Technology*, 26. DOI: 10.1007/s10676-024-09757-7.

[38] E. Iglesias, S. Jozashoori, and M. Vidal. 2023. Scaling up knowledge graph creation to large and heterogeneous data sources. *J. Web Semant.*, 75, 100755. DOI: 10.1016/j.websem.2022.100755.

[39] E. Jimenez-Ruiz and B. Cuenca Grau. 2011. Logmap: logic-based and scalable ontology matching. In *The Semantic Web – ISWC 2011*. L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. Noy, and E. Blomqvist, (Eds.) ISBN: 978-3-642-25073-6.

[40] A. Jobin, M. Ienca, and E. Vayena. 2019. The global landscape of AI ethics guidelines. en. *Nature Machine Intelligence*, 1, 9, (Sept. 2019), 389–399. DOI: 10.1038/s42256-019-0088-2.

[41] A. M. Kaushik and R. Mutharaju. 2021. Chapter 21. an ontology design pattern for modeling bias. In *Studies on the Semantic Web*. IOS Press, (May 2021). DOI: 10.3233/ssw210024.

[42] E. F. Kendall and D. L. McGuinness. 2019. *Ontology Engineering. Synthesis Lectures on the Semantic Web: Theory and Technology*. Morgan & Claypool Publishers. DOI: 10.2200/S00834ED1V01Y201802WBE018.

[43] A. A. Khan, S. Badshah, P. Liang, et al. 2022. Ethics of ai: a systematic literature review of principles and challenges. In *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering* (EASE '22). Association for Computing Machinery, Gothenburg, Sweden, 383–392. ISBN: 9781450396134. DOI: 10.1145/3530019.3531329.

[44] T. Lebo, S. Sahoo, D. McGuinness, et al. 2013. *PROV-O: The PROV Ontology*. English. *W3C Recommendation*. World Wide Web Consortium, United States, (Apr. 2013).

[45] H. Li, G. Appleby, C. D. Brumar, et al. 2024. Knowledge graphs in practice: characterizing their users, challenges, and visualization opportunities. *IEEE Transactions on Visualization and Computer Graphics*, 30, 1, 584–594. DOI: 10.1109/TVCG.2023.3326904.

[46] W. Mark and et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3, 1, 1–9.

[47] N. Mehrabi, F. Morstatter, N. Saxena, et al. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54, 6, Article 115, (July 2021), 35 pages. DOI: 10.1145/3457607.

[48] N. Mehrabi, F. Morstatter, N. A. Saxena, et al. 2019. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54, 1–35.

[49] M. Miceli, T. Yang, L. Naudts, et al. 2021. Documenting computer vision datasets: an invitation to reflexive data practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21). ACM, 161–172. ISBN: 9781450383097. https://doi.org/10.1145/3442188.3445880.

[50] A. J. Miles and S. Bechhofer. 2009. Skos simple knowledge organization system reference. In *Simple Knowledge Organization System*. https://api.semanticscholar.org/CorpusID:58835891.

[51] M. Mitchell, S. Wu, A. Zaldivar, et al. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (FAT* '19). ACM, 220–229. ISBN: 9781450361255. DOI: 10.1145/3287560.3287596.

[52] S. U. Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, New York.

[53] N. Noy and C. Goble. 2022. Are we cobblers without shoes? making computer science data fair. *Commun. ACM*, 66, 1, (Dec. 2022), 36–38. DOI: 10.1145/3528574.

[54] A. Olteanu, C. Castillo, F. Diaz, et al. 2019. Social data: biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2. DOI: 10.3389/fdata.2019.00013.

[55] E. Parliament and C. of the European Union. 2021. *Proposal for a Regulation laying down harmonised rules on artificial intelligence*. European Commission.

[56] E. Pitoura. 2020. Social-minded measures of data quality: fairness, diversity, and lack of bias. *J. Data and Information Quality*, 12, 3, Article 12, (July 2020), 8 pages. DOI: 10.1145/3404193.

[57] M. Poveda-Villalon, A. Gomez-Perez, and M. C. Suarez-Figueroa. 2014. OOPS! (OntOlogy Pitfall Scanner!): An On-line Tool for Ontology Evaluation. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 10, 2, 7–34.

[58] G. C. Publio, D. Esteves, A. Ławrynowicz, et al. 2018. Ml-schema: exposing the semantics of machine learning with schemas and ontologies. (2018). DOI: 10.48550/ARXIV.1807.05351.

[59] I. D. Raji, I. E. Kumar, A. Horowitz, et al. 2022. The fallacy of ai functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '22). Association for Computing Machinery, Seoul, Republic of Korea, 959–972. ISBN: 9781450393522. DOI: 10.1145/3531146.3533158.

[60] I. D. Raji, A. Smart, R. N. White, et al. 2020. Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (FAT* '20). ACM, Barcelona, Spain, 33–44. ISBN: 9781450369367. DOI: 10.1145/3351095.3372873.

[61] I. D. Raji and J. Yang. 2019. About ml: annotation and benchmarking on understanding and transparency of machine learning lifecycles. *ArXiv*, abs/1912.06166.

[62] C. Reddy, D. Sharma, S. Mehri, et al. 2021. Benchmarking bias mitigation algorithms in representation learning through fairness metrics. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/2723d092b63885e0d7c260cc007e8b9d-Paper-round1.pdf.

[63] G. P. +. A. Research. 2021. Know your data. https://knowyourdata.withgoogle.com/docs/. Access:10.06.2022. (2021).

[64] H. F. Research. 2022. Data Measurements Too. https://huggingface.co/spaces/huggingface/data-measurements-tool. (2022).

[65] P. Reyero-Lobo, E. Daga, H. Alani, et al. 2023. Semantic web technologies and bias in artificial intelligence: A systematic literature review. *Semantic Web*, 14, 4, 745–770. DOI: 10.3233/SW-223041.
[66] P. Reyero-Lobo, J. Kwarteng, M. Russo, et al. 2024. A multidisciplinary lens of bias in hate speech. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining* (ASONAM '23). Association for Computing Machinery, Kusadasi, Turkiye, 121–125. ISBN: 9798400704093. DOI: 10.1145/3625007.3627491.
[67] J. Riley. 2021. The elusive promise of ai: a second look. *Ubiquity*, 2021, April, Article 1, (Apr. 2021), 10 pages. DOI: 10.1145/3458742.
[68] M. Russo, Y. Chudasama, D. Purohit, et al. 2024. Employing hybrid ai systems to trace and document bias in ml pipelines. *IEEE Access*, 12, 96821–96847. DOI: 10.1109/ACCESS.2024.3427388.
[69] M. Russo and M.-E. Vidal. 2024. Leveraging ontologies to document bias in data. In *Proceedings of the Second Workshop on Fairness and Bias in AI* (CEUR Workshop Proceedings) number 3808 (Santiago de Compostela, Spain, Oct. 20, 2024). R. Calegari, V. Dignum, and B. O'Sullivan, (Eds.) Aachen, 10–21. https://ceur-ws.org/Vol-3808/paper5.pdf.
[70] A. Scherp, G. Groener, P. Skoda, et al. 2024. Semantic Web: Past, Present, and Future. *Transactions on Graph Data and Knowledge*, 2, 1, 3:1–3:37. DOI: 10.4230/TGDK.2.1.3.
[71] R. Schwartz, A. Vassilev, K. K. Greene, et al. 2022. Towards a standard for identifying and managing bias in artificial intelligence. en. (Mar. 2022). DOI: https://doi.org/10.6028/NIST.SP.1270.
[72] N. Shahbazi, Y. Lin, A. Asudeh, et al. 2023. Representation bias in data: a survey on identification and resolution techniques. *ACM Comput. Surv.*, 55, 13s, Article 293, (July 2023), 39 pages. DOI: 10.1145/3588433.
[73] R. Shelby, S. Rismani, K. Henne, et al. 2023. Sociotechnical harms of algorithmic systems: scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '23). Association for Computing Machinery, Montréal, Canada, 723–741. ISBN: 9798400702310. DOI: 10.1145/3600211.3604673.
[74] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. 2017. Fake news detection on social media: a data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19, 1, 22–36.
[75] S. Srinivasan, C. Dickens, Augustine, et al. 2022. A taxonomy of weight learning methods for statistical relational learning. *Mach. Learn.*, 111, 8, (Aug. 2022), 2799–2838. DOI: 10.1007/s10994-021-06069-5.
[76] E. Stamboliev and T. Christiaens. 2024. How empty is trustworthy ai? a discourse analysis of the ethics guidelines of trustworthy ai. *Critical Policy Studies*, 0, 0, 1–18. eprint: https://doi.org/10.1080/19460171.2024.2315431. DOI: 10.1080/19460171.2024.2315431.
[77] J. Stoyanovich, S. Abiteboul, B. Howe, et al. 2022. Responsible data management. *Commun. ACM*, 65, 6, (May 2022), 64–74. DOI: 10.1145/3488717.
[78] C. Sun, A. Asudeh, H. V. Jagadish, et al. 2019. Mithralabel: flexible dataset nutritional labels for responsible data science. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (CIKM '19). Beijing, China, 2893–2896. ISBN: 9781450369763. DOI: 10.1145/3357384.3357853.
[79] H. Suresh and J. Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization* (EAAMO '21) Article 17. –, NY, USA, 9 pages. ISBN: 9781450385534. DOI: 10.1145/3465416.3483305.
[80] M. van Bekkum, M. de Boer, F. van Harmelen, et al. 2021. Modular design patterns for hybrid learning and reasoning systems: a taxonomy, patterns and use cases. *Applied Intelligence*, 51, 9, (Sept. 2021), 6528–6546. DOI: 10.1007/s10489-021-02394-3.
[81] M.-E. Vidal, K. M. Endris, S. Jazashoori, et al. 2019. Transforming Heterogeneous Data into Knowledge for Personalized Treatments—A Use Case. en. *Datenbank Spektrum*, 19, 2, (July 2019), 95–106. DOI: 10.1007/s13222-019-00312-z.
[82] A. Wang, A. Liu, R. Zhang, et al. 2022. Revise: a tool for measuring and mitigating bias in visual datasets. *Int. J. Comput. Vision*, 130, 7, (July 2022), 1790–1810. DOI: 10.1007/s11263-022-01625-5.
[83] J. Whittlestone, R. Nyrup, A. Alexandrova, et al. 2019. The role and limits of principles in ai ethics: towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '19). Association for Computing Machinery, Honolulu, HI, USA, 195–200. ISBN: 9781450363242. DOI: 10.1145/3306618.3314289.
[84] L. Yu. 2011. Foaf: friend of a friend. In *The friend of a friend(foaf) project*. https://api.semanticscholar.org/CorpusID:60893017.

A Doc-BiasO Axiomatization

The conceptualization of the Doc-BiasO ontology is specified using OWL logical axioms. This allows for consistency checks and logical inferences on a resulting RDF knowledge graph. In the following, we exemplify some domain range axioms for the `Bias` and `Bias Evaluation` classes, as well as axioms denoting restrictions on Bias. As already mentioned, OWL is formally defined in description logics, thus first, we summarize in Table 8, the relevant Description Logic operators and their descriptions.

Table 8. **Operators.** Conventional Notation of Description Logic Operators.

Symbol	Description
$\sqsubseteq$	Concept inclusion
$\forall$	Universal restriction
$\exists$	Existential restriction

Bias.

$$Bias \sqsubseteq \forall hasBiasMeasure.BiasMeasure \quad (Domain)$$
$$\exists hasBiasMeasure.BiasMeasure \sqsubseteq Bias \quad (Range)$$

$$Bias \sqsubseteq \forall isAssociatedTo.Application \quad (Domain)$$
$$\exists isAssociatedTo.Application \sqsubseteq Bias \quad (Range)$$

$$Bias \sqsubseteq \forall isAlignedWith.Harm \quad (Domain)$$
$$\exists isAlignedWith.Harm \sqsubseteq Bias \quad (Range)$$

$$hasBiasMeasure\ some\ BiasMeasure$$
$$isAlignedWith\ some\ Harm$$
$$isAssociatedTo\ some\ Application$$

Class disjointness between all four classes is stated.

BiasEvaluation.

$$BiasEvaluation \sqsubseteq \forall evaluatesWith.BiasMeasure \quad (Domain)$$
$$\exists evaluatesWith.BiasMeasure \sqsubseteq BiasEvaluation \quad (Range)$$

$$BiasEvaluation \sqsubseteq \forall wasAttributedTo.Document \quad (Domain)$$
$$\exists wasAttributedTo.SDocument \sqsubseteq BiasEvaluation \quad (Range)$$

$$BiasEvaluation \sqsubseteq \forall evaluatesIn.Dataset \quad (Domain)$$
$$\exists evaluatesIn.Dataset \sqsubseteq BiasEvaluation \quad (Range)$$

$$BiasEvaluation \sqsubseteq \forall evaluatesFor.Task \quad (Domain)$$
$$\exists evaluatesFor.Task \sqsubseteq BiasEvaluation \quad (Range)$$

Received 3 June 2025; revised 18 July 2025; accepted 18 July 2025