An Ontology for Fairness Metrics

Jade S. Franklin  
Rensselaer Polytechnic Institute  
Troy, New York, USA

Karan Bhanot  
Renssеляer Polytechnic Institute  
Troy, New York, USA

Mohamed Ghalwash  
IBM Research  
Yorktown, New York, USA

Kristin P. Bennett  
Rensselaer Polytechnic Institute  
Troy, New York, USA

Jamie McCusker  
Rensselaer Polytechnic Institute  
Troy, New York, USA

Deborah L. McGuinness  
Rensselaer Polytechnic Institute  
Troy, New York, USA

ABSTRACT
Recent research has revealed that many machine-learning models and the datasets they are trained on suffer from various forms of bias, and a large number of different fairness metrics have been created to measure this bias. However, determining which metrics to use, as well as interpreting their results, is difficult for a non-expert due to a lack of clear guidance and issues of ambiguity or alternate naming schemes between different research papers. To address this knowledge gap, we present the Fairness Metrics Ontology (FMO), a comprehensive and extensible knowledge resource that defines each fairness metric, describes their use cases, and details the relationships between them. We include additional concepts related to fairness and machine learning models, enabling the representation of specific fairness information within a resource description framework (RDF) knowledge graph. We evaluate the ontology by examining the process of how reasoning-based queries to the ontology were used to guide the fairness metric-based evaluation of a synthetic data model.

CCS CONCEPTS

KEYWORDS
fairness metric, bias, RDF knowledge graph, machine learning evaluation

ACM Reference Format:

INTRODUCTION
With the growing ubiquity of machine learning (ML) models in corporate, governmental, and academic environments, the decisions of ML models now more than ever have the potential to change lives. ML models filter job applications, detect disease, and even predict whether convicts are likely to reoffend, to name just a few examples. However, there is a growing concern that many of these ML models are trained on biased data, or otherwise make unfair predictions about minority groups. As an example, a 2016 news report generated controversy when it alleged that a recidivism-prediction algorithm in use by a Florida courthouse was biased against people of color [1].

Although recognizing when models are biased is highly important, it can be difficult to actually measure unfairness. Since there is no single definition of fairness, researchers have instead created a wide variety of fairness notions: definitions of fairness that a given model can either satisfy or fail to satisfy [25], as well as fairness metrics: mathematical formulae for measuring how close a given model is to satisfying some fairness notion [16] (for example, “Equal Opportunity” is a fairness notion, and “Equal Opportunity Difference” is a fairness metric that measures it). Furthermore, not all metrics can be satisfied in most cases, so understanding the relationships between different metrics—and choosing the correct one—is vital. However, a lack of consistent naming, ambiguity between fairness metrics and notions, and the wide variety of literature published on the topic can cause confusion for ML model developers on how to actually use these metrics appropriately [30].

To aid in the development of fair ML models, and provide guidance to researchers on how best to utilize fairness notions and metrics, we present the Fairness Metrics Ontology (FMO). FMO is a comprehensive and extensible knowledge resource that defines each fairness notion and metric, describes their use cases, and details the relationships between them. FMO also includes related concepts for describing ML models, such as the statistical metrics used to evaluate them, and defines how each fairness notion can be derived from these metrics. Although FMO can be used as a standalone resource by itself, it can also be used to model fairness information for specific ML models in a Resource Description Framework (RDF) knowledge graph, and can be combined with a reasoner to make inferences about which fairness metrics a model satisfies.

We note that FMO was designed from the perspective of formal fairness research, focusing on mathematical definitions of fairness and analysis of their results, and leave ethical considerations of which fairness approaches are better to the user. Instead, we designed FMO to provide metrics based on a user’s individual values and use case, and avoid making judgements on our personal opinions by grounding the ontology in the formal fairness literature. We describe herein relevant background on fairness metrics, fairness
notions, and ontologies (Section 2), related work (Section 3), the structure and layout of the ontology itself (Section 4), and a case study of a scenario in which the ontology was used to recommend fairness metrics (Section 5).

2 BACKGROUND

Research into formal definitions of fairness has grown tremendously in recent years, and so have the number and variety of fairness metrics. Using these fairness metrics, researchers can test how well their models conform to specific definitions of fairness. However, many of the prominent ML-based systems suffer from multiple forms of bias and unfairness in their predictions, particularly with regards to race and gender, and in many cases the unfairness is not noticed until the models are already operating in the real world, e.g., the inability of facial recognition software to detect black, female faces [8, 9], bias against female candidates in recruitment algorithms [12, 14], and saliency cropping algorithms used by Twitter, Google and Apple under-prioritizing non-white, non-male faces [7]. A survey by Kiritchenko and Mohammad [24] found statistically significant bias in the majority of the 200 sentiment analysis models evaluated.

Although a great deal of information on fairness metrics has been published in the form of literature reviews, we find that the complexity and sheer number of metrics can overwhelm researchers and cause confusion in how best to select, understand, and interpret one or more metrics. We therefore propose that our fairness metrics ontology can act as an accessible knowledge resource to guide data scientists in using fairness metrics, and hypothesize that the use of this ontology will allow scientists to better measure, evaluate, and improve the fairness of their ML models.

2.1 Distinguishing Fairness Metrics

Fairness metrics are used to discern information about how fairly a ML model or dataset treats individuals from different groups. Due to the complexity of assessing fairness, there is not a single fairness metric that can be prioritized over all others. Instead, as the three major categories of fairness notions (Independence, Separation, and Sufficiency) cannot in general be satisfied for the same situation, one has to choose which notion(s) of fairness most closely match the needs of the application [2]. Furthermore, the developer must decide whether to use a stricter or more relaxed fairness notion, and should also determine whether or not he or she needs to account for biases such as intersectionality and masking, whether group or individual fairness should be prioritized, and whether the metric itself will use ratios, difference, or a statistical test to measure the chosen fairness notion. Once the metric itself has been chosen and evaluated, the researcher must also interpret the results of the metric, and determine whether the measurement meets the appropriate threshold to state whether or not a model or dataset is fair—according to the measured fairness notion. This can be very difficult for a non-expert audience to understand; a 2020 study by Saha et al. [30] showed participants had difficult distinguishing between metrics, and in particular struggled to understand how the Equal Opportunity metric works. While there exist some resources for assisting with using and understanding fairness metrics, as described in Section 3, the task grows in complexity when comparing multiple metrics against one another. For this reason, we propose that an ontology will provide a thorough and consistent knowledge resource to aid in representation and understanding of ML evaluation as it pertains to fairness.

2.2 Ontologies and Semantics

Ontologies provide a consistent, interoperable and machine-readable method of structuring knowledge, and provide the structure of knowledge that enables data to be stored in knowledge graphs. Both ontologies and knowledge graphs are represented as triples in \((subject, predicate, object)\) form in a property graph where \(subject\) and \(object\) are nodes in the graph with an edge labeled with a \(predicate\) (or property) between them. If these graphs are connected to the internet, they serve as a form of semantic, linked data. Via the network of ontologies and knowledge graphs that make up the Semantic Web, data is made semantically linked—linked to web resources which unambiguously define its meaning—and represented in a consistent and broadly compatible way. There are multitudes of benefits to storing data semantically, and we outline three main pillars of semantic technology—ontologies, knowledge graphs, and reasoning—below. In Section 4, we describe how the Fairness Metrics Ontology makes use of these three technologies to guide developers when evaluating and understanding the fairness of their machine-learning models.

2.2.1 Ontologies. Structuring knowledge in an ontology has several advantages. At the most basic level, an ontology is used to collect information about a topic, serving as a single, self-consistent resource that assigns unique, unambiguous identifiers (uniform resource identifiers, or URIs) to concepts, defines and describes them, and organizes them into a class hierarchy. By browsing the ontology manually, a user can look up class definitions (along with other properties the ontology includes, such as synonyms, examples of usage, and sources of this information), and view the subclass/superclass hierarchy trees to understand how the different concepts fit together. Since an ontology assigns unique URIs to each concept, a data scientist can use these URIs to describe data unambiguously. The ontology is also machine-readable, and can be traversed by a software agent like any other property graph. The Web Ontology Language, or OWL [3], provides a single language with which to design cross-compatible ontologies. OWL also supports property restrictions—these include value constraints, which add restrictions on what classes can be the domain or range of a property, and cardinality constraints, which control the number of values a class’s property may refer to. Using these constraints (as well as subclass/superclass relationships) enables reasoning, including reasoning over knowledge graphs.

2.2.2 Knowledge Graphs (KGs). KGs provide a way to store data in a property graph that instantiates the concepts defined in an ontology. If an ontology defines the class “Person,” a KG could describe a specific individual that is an instance of the Person class—and if the ontology defines “Person” as a subclass of “Agent,” any specific person would by extension also be considered an Agent. KGs are frequently represented with the Resource Description Framework (RDF) [13], and can be queried using the SPARQL Protocol and RDF Query Language [19]. As KGs are graph structures, querying
them takes advantage of graph patterns and algorithms in order to traverse a specific path of data, and as they are semantic, the identifiers for classes, instances, and predicates can be resolved as hyperlinks to RDF documents online that provide definitions, context, and other metadata. In addition, they are fairly scalable, as they can be easily updated, and can be augmented with provenance encodings to track metadata about when information was uploaded, what its sources are, etc. Additionally, the underlying ontology’s property restrictions and class hierarchy can be taken advantage of to perform reasoning over a KG.

2.2.3 Reasoning. Reasoning is a method of inferring new facts from existing facts according to a specific set of rules, such as those defined by an ontology. Using subclass-superclass relations, a reasoner infers which classes a given instance belongs to, and using property restrictions, types and other information about the subject and object of a given property are inferred. By making several inferences over an entire ontology or KG, a reasoner can deduce detailed information about the relationships between different represented entities, and can even infer the existence of new entities according to what is necessary to fulfill property restrictions.

3 RELATED WORK

Prior work has recognized the need for robust tools and guidance to assist with the development of fair ML models, and we build off of the work of other authors who have made strides in this direction. In particular, we note that several authors have conducted exhaustive literature reviews of the fairness metrics field, and their contributions have enabled us to develop a thorough set of metrics to use within our ontology. Verma and Rubin [32] provide a highly explanatory and practicality-oriented review of fairness metrics in the field. Caton and Haas [10] provide a thorough review of fairness metrics, and categorize these metrics with an in-depth categorization scheme. Garg et al. [16] analyzed the contextual, logical, and mathematical relationships between fairness metrics in their review. Mehrabi et al. [27] surveyed the area of biases as well as fairness metrics, and made some steps in the direction of connected biases to the fairness metrics that measure them. Makhlof et al. [25] formalized a distinction between fairness notions and fairness metrics, and provided additional guidance on when best to use each metric in the form of a flow chart. Lastly, Hutchinson and Mitchell [20] provide a thorough overview of the developments and changes in fairness research over the past fifty years, and track how understanding and terminology have evolved over time.

Of these reviews provides a significantly broad and in-depth survey of fairness metrics, although we note that the sheer number of metrics, as well as the several different variables that come into play during metric selection, present a challenge for researchers not intimately familiar with the field. Additionally, the terminology is not always consistent between reviews—metrics frequently have several different names, and in some cases may be defined slightly differently. However, more approachable tools for beginners do exist. The AI Fairness 360 toolkit (AIF360), an open-source toolkit for analysis, report, and mitigation of bias in ML models [4], provides a useful and beginner-friendly toolkit for aiding ML model developers. The documentation for the toolkit is thorough, and flowchart-based tutorials guide users in choosing a fairness metric, and the toolkit itself claims to support over seventy different fairness metrics. However, AIF360 does not include machine-readable definitions, formulas, and other information about fairness metrics, and its analyses do not use information about the relationships between fairness metrics to infer information about metrics which imply other metrics, or metrics which are mutually incompatible. Additionally, as AIF360 does not include any sort of ontology, AIF360 cannot be used to flexibly structure information about fairness metrics, such as by tracking fairness metrics and results for different versions of a model in a KG. In terms of other non-ontology fairness analysis tools, Kim et al. [23] developed the FACT (FAirness-Confusion Tensor) diagnostic specifically to analyze trade-offs between different fairness metrics. Similarly to our evaluation of FMO, Kim et al. evaluate FACT by demonstrating its results when applied to a synthetic data fairness case study.

On the other hand, Panigutti et al. [28] introduced “Doctor XAI,” an ontology-enabled method for generating explanations for black-box machine learning models. The approach provides an excellent example of how knowledge-based methods can enable improved ML model understanding. We also note that a general-purpose metrics ontology was presented by Soergel and Helfer [31] in 2016, which provides a platform for knowledge-based representation and analysis of program measurement, and argues that the proper definition and documentation of what exactly metrics are measuring enables increased understanding and clarity during the analysis process. However, the general-purpose metrics ontology was not developed for application towards ML model evaluation (it is instead primarily focused on measurement of clinical programs or for medical research) and does not contain metrics useful for ML model fairness evaluation. This prior work demonstrates the utility of an ontology for defining any kind of metric, including fairness metrics.

4 METHODS

In order to provide a useful knowledge resource for fairness metric representation, we needed to ensure FMO was an accurate and fully-featured ontology of fairness metrics that is as complete as possible. Given the several challenges of selecting and interpreting fairness metrics and definitions described in Section 2.1, we developed FMO to represent all information about fairness metrics relevant to these challenges and that would be useful for a researcher or automated process to have easy access to.

We define three main goals for FMO:

1. FMO should thoroughly and accurately define and describe a complete set of fairness metrics and their related concepts and properties.
2. FMO should be usable as a structure for instantiating ML model evaluation data in KGs.
3. Via reasoning, FMO should be able to make automated inferences that aid in selecting and interpreting fairness metrics and fairness-based ML evaluations.

In order to meet the first goal, we decided on an initial category of metrics (group fairness metrics) that was commonly used and well-represented in the literature, and gathered the data needed to determine definitions and other details for each metric (Section 4.1). To meet the second goal, we built an ecosystem of additional
concepts and definitions related to fairness metrics, and determined appropriate relationships between them needed for ontology-based modeling (Section 4.2). To meet the third goal, we developed queries to test the use of reasoning with the ontology, and detail a case study on reasoning with the ontology to recommend metrics (Section 5.1).

4.1 Design Process

In order to gather the appropriate amount of background information, we reviewed several literature surveys of the fairness metrics field, as well as the original fairness metric papers these surveys cited. In some cases, there were discrepancies involving different papers using different names for different metrics, or even defining the same metric differently; in these cases, we referred back to how the original paper defined the metric, and found that in general errors of incorrectly defining a metric were uncommon and not widely proliferated. Determining the name to use for a metric was less straightforward, as several metrics had a large number of names that varied across several papers, so for these metrics we selected the most common name that remained consistent with other metric names and recorded the others as alternate names in the ontology. We kept track of the provenance of the knowledge collected during this process, such that individual facts are linked back to the paper(s) where that information was acquired from.

The process of designing the ontology required additional synthesis beyond just copying over names and definitions, and the first of these was resolving the ambiguity between fairness definitions and metrics. Frequently, the same term is used to describe both a fairness property which a model may or may not satisfy depending on whether it meets certain conditions, as well as the fairness metric which measures the degree to which a model does or does not satisfy a property. To resolve this ambiguity, we chose the term “Fairness Notion,” as used in Makhlouf et al. [25] to describe a specific fairness definition or property which a model may or may not satisfy, and used “Fairness Metric” to refer to a method of measuring the degree to which a model conforms to some fairness notion. By resolving this ambiguity, FMO is able to capture the different methods used to measure fairness (relative difference, ratio, and so on) as their own Fairness Metric concepts distinct from the fairness notions they measure. Using property restrictions, we ensure that the structure of the ontology is properly represented: using cardinality and domain and range constraints, it is ensured that each instance of a fairness metric will always measure exactly one fairness notion.

In order to ensure that the ontology was as accurate as possible, we needed to develop class hierarchies that both corresponded to common categorization schemes and were logically sound according to the Liskov substitution principle. For example, it is extremely common to group statistically-derived fairness notions into the categories corresponding to the Independence, Separation, and Sufficiency notion, but not every notion in these categories is a strict subclass of the notion it is grouped under. For example, the Equal Opportunity notion is a more relaxed version of Separation, so although Equal Opportunity is categorized within the Separation group, a model which satisfies Equal Opportunity does not necessarily satisfy Separation. In contrast, Separation always implies Equal Opportunity, so Separation is actually a subclass of Equal Opportunity, not the other way around. In order to represent the idea that Separation, Equal Opportunity, and other metrics all fall under the Separation category, we defined the broader concept “Separation Class Fairness Notion” to encompass all fairness notions derived from this notion, and to avoid confusion used the name “Equalized Odds” as the primary term for the fairness notion also known as Separation. Encountering unintuitive relationships between metrics in this way was extremely common, and reducing the ambiguity by providing a single, consistent set of fairness notions and relationships between them was an important goal of FMO.
Figure 2: Fairness Notions, as represented in the Fairness Metrics Ontology. On the left is the class hierarchy of group fairness notions, while on the right are the relaxed fairness notions and biases. Some classes have been omitted for space.

4.2 Representation

FMO is modeled in OWL using the Semanticscience Integrated Ontology (SIO) [15] as an upper level ontology, providing basic properties such as hasAttribute, inRelationTo, isDerivedFrom, and so on. To model some statistical metrics, such as precision, recall, and so on, we use the Statistics Ontology (STATO) [17]. Building off of these ontologies, FMO captures three main types of concepts: fairness notions, evaluation metrics, and ML concepts. We present how these three groups fit together to make up the ontology in the high-level concept map shown in Figure 1.

4.2.1 Fairness Notions. The major component of the ontology is its fairness notions: the specific properties of fairness that an ML model has as an attribute if it satisfies certain requirements [25] (See Figure 2). We provide both formal, probability-based definitions of fairness, as well as human-friendly descriptions that relate them to the statistical metrics they correspond to; for example, we define statistical parity as satisfied if the probability of a model predicting a good outcome or a bad outcome for an individual is independent of whether the individual is a member of a protected group, but we describe statistical parity as resulting when each protected class has the same positive rate. In addition, for each fairness notion we define alternate terms, e.g. Statistical Parity is also known as Independence, and also include an example of usage: a description of the situation under which prioritizing a fairness notion would be useful—for Statistical Parity, this is any situation in which emphasizing equality among groups is prioritized without accounting for mitigating factors that might cause one group to under-perform.

Additionally, we define biases [27]—different types of unfairness—that the notion of fairness is well-suited to detect. Returning to the previous example, measurements of statistical parity can detect historical bias—lower performance due to historical prejudice against a group—and prioritizing statistical parity in a model ensures that the model selects a positive outcome to each group at the same rates. However, there are certain pitfalls with this scenario. Just because statistical parity can measure historical bias does not mean a lack of statistical parity implies historical bias—as there are an unbounded number of biases that any fairness notion can be sensitive to, it is perfectly possible that a lack of statistical parity is the result of some other form of bias, such as population bias. Secondly, while statistical parity may be able to mitigate historical bias, it is highly vulnerable to masking [25]—situations where unfairness is deliberately hidden in order to meet certain requirements—but other forms of inequality remain. This could take the form of reduced accuracy in giving positive outcomes to members of the protected group who deserve it, or greater leeway given to members of the protected group—two other kinds of biases that are better addressed by fairness notions such as equalized odds. Determining biases that fairness notions can measure, as well as determining which biases fairness notions are poor at measuring, is something of a subjective process, and there was not a great deal of research into this area, but enough information was provided by the surveys by Mehrabi et al. [27] and Makhlouf et al. [25] that we felt comfortable creating a framework of notion-bias relations that should aid users of FMO in determining the correct fairness notion for a given situation.

We also model the relationships between notions: circumstances in which one notion may be inferred from another, or sets of mutually exclusive notions. These are defined via the subclass-superclass relations as mentioned in Section 4.1, such that if a model satisfies one notion, it will also satisfy all superclasses of that notion. In
terms of determining whether or not a model actually satisfies a notion, it is most common to use a relaxed version of a fairness notion—for this purpose, we provide the Relaxed Fairness Notion superclass, which is derived from any other fairness notion, and has two subclasses: the Thresholded Fairness Notion, which is associated with a threshold that must be met in order for the notion to be satisfied, and the Statistical Test-Based Fairness Notion, which is satisfied when a statistical test for fairness resolves with a specific p-value. By measuring these relaxed fairness notions, the computed result of a fairness Metric can imply whether or not a model satisfies a fairness notion (see Figure 3). Group fairness notions are identified in the ontology as being derived from specific statistical metrics, as they are satisfied if the value of these statistical metrics is the same for each protected group. As an example, statistical parity is derived from positive rate, so statistical parity is satisfied if each group is given the positive outcome at the same rate.

As fairness notions themselves are Boolean properties—a model either satisfies them, or it does not—fairness metrics serve the purpose of providing a continuous value measurement of the degree to which a fairness notion is or is not satisfied. We define difference-based and ratio-based fairness metrics, which measure the difference or ratio of statistical metrics between different groups. Again, in the case of statistical parity, the difference or ratio between one group and others in positive rate is the result of the difference-based fairness metric and ratio-based fairness metric, respectively. If a thresholded fairness notion is measured, the notion is satisfied if the value of the metric is at or below the threshold for each group (e.g. a difference of 0.1 in positive rate between groups means a thresholded fairness notion with a threshold of 0.1 or more would be satisfied).

Additionally, we define the statistical fairness test, which uses the probability-based nature of the notions to determine the likelihood that a given fairness metric is satisfied via calculating p-values. As statistical parity is defined as an equal probability of members of each group getting a positive outcome, the statistical fairness test can be used to estimate the chances that the observed positive rate between groups reflects an equal probability of the positive outcome being assigned. Using the Statistical Test-Based Fairness Notion, the underlying fairness notion is considered to be satisfied if the p-value of the statistical fairness test is at or below a certain value.

The evaluation metrics are intended to be used to represent data about ML models in a KG, where multiple instances of statistical metrics, fairness metrics, and fairness notions can all be represented in different configurations and implying different inferences. In this way, these classes can be used to represent the results of a specific ML evaluation, according to the results of a specific ML model with a specific testing dataset. To assist with the representation of ML models in a KG, we define specific ML concepts to have notions and metrics as properties.

### 4.2.3 Machine Learning Concepts
FMO defines several concepts related to ML models and evaluation, in order to facilitate the modeling of information about ML models and fairness in a KG (See Figure 4). The ML model itself is a class, and has its evaluation metrics and any fairness notions it satisfies as attributes. The model evaluation itself, (essentially just the confusion matrix of results from comparing an ML model’s output to the ground truth of a testing dataset, takes the machine learning model and any number of dataset cohorts as input, and outputs evaluation metrics. The evaluation metrics record results for the testing dataset as a whole, as well as the individual cohorts of a dataset: subgroups of race, sex, etc. which may be protected and which are compared to determine fairness.

Essentially, an ML model may have any number of evaluations, each of which may be performed on any set of dataset cohorts, and will produce any number of evaluation metrics. As the metrics...
themselves are represented in relation to individual cohorts, an automated reasoner can compare the evaluation metric results in order to determine which, if any, fairness notions are satisfied. In this way, we have designed FMO to support modeling a complete KG that supports an arbitrary number of models, datasets, and evaluations, and can be used as a basis for in-depth analysis of fairness information.

5 RESULTS AND DISCUSSION

We present FMO, a knowledge infrastructure to guide researchers on selecting, interpreting, and understanding fairness metrics and ML model evaluations. The ontology can be used as a knowledge resource directly, and also supports usage as a basis for KG representation and reasoning about fairness metrics.

Since FMO represents the biases that are typically addressed or not addressed by fairness notions, as well as the underlying statistical metrics which are used to calculate fairness metrics, FMO can guide a user in the selection of fairness notions and metrics that meet specific requirements. As an example, a user may need to select a metric in order to measure the fairness of a model that predicts whether a given patient is at risk of developing diabetes, and will train and test an ML model on a specific dataset. By comparing the relative count of each cohort against public data on the demographics of diabetes patients, the user identifies that the dataset suffers from representation bias: the dataset has a lower number of Hispanic or Latino subjects than in the population at large. Additionally, as the ML model is intended for a medical setting, the user decides the model should prioritize reducing the number of false negatives.

Using FMO, the user can query for fairness notions that are both 1) able to address representation bias, and 2) prioritize reducing the number of false negatives (see Figure 5). In this instance, FMO returns equal opportunity (which prioritizes the measurement of false negatives alone), equalized odds (which prioritizes both false positives and false negatives), and balance for positive class (which is a specialized version of equal opportunity), and does not return notions such as total fairness (which cannot be satisfied in the case of representation bias). As this is a description logic query, the explanation for the reasoning is displayed to the user, and the user can further refine as necessary.

5.1 Case Study: Synthetic Data Fairness

Artificial intelligence (AI) applications require access to vast quantities of data in order to make accurate generalizations in real-world scenarios. Properly trained AI applications have proven themselves significantly useful in the healthcare domain; however, access to the patient data required to train these applications is often restricted by privacy laws. As an alternative, researchers have introduced synthetic data: artificially-generated data acting as a proxy for real data without exposing sensitive patient information. While several data generators have been developed, HealthGAN [33, 34] has been shown to be effective in generating high-quality healthcare data for both public and private datasets and thus, will be used for data generated in our analysis of synthetic data fairness.

With even real datasets often being biased, it is doubly important to evaluate the fairness of synthetic datasets before they are used in the real world—otherwise, it is possible for biases in the original data to be propagated and amplified in the corresponding synthetic data. In Kenfack et al. [22], the authors measured biases by observing the distributions of the generated image data and found that synthetic data generation favored one class over the other. In Bhanot et al. [6], the authors compared the synthetic data with the real data using log disparate impact metric for temporal and non-temporal data, inspired by time-series metrics for synthetic data [5]. They found
Figure 5: **Top left:** A description logic query to return fairness notions which are derived from false negative measurements. **Top right:** A query similar to the left which also requires metrics to measure representation bias. **Bottom:** An explanation for the “equalized odds” fairness notion (for answering the top right query).

that synthetic data over-represents some and under-represents other subgroups of protected attributes. In Cheng et al. [11], the authors generated ML models on the synthetic data and found them favoring certain classes over others using the Cook’s distance. Alternatively, if unfairness is discovered and accounted for, it is possible to generate synthetic data that is more fair than the original data. Gupta et al. [18] found the synthetic data generated using Generative Adversarial Networks (GANs) except PATE-GAN to improve fairness with regards to specific fairness notions (statistical parity, equal opportunity, and equalized odds).

These results suggest that there are numerous ways to measure biases in synthetic data, often defined by the researcher and the problem. However, the choice of which fairness metric to choose can be dependent on the researcher or domain and might not be objective: Yeom and Tschantz [35] showed inherent trade-offs between different definitions of fairness make choosing the correct metric as much a question of worldview as of math, and Kim et al. [23] demonstrate that these trade-offs are significant when analyzing synthetic data fairness. As a demonstration and case study of its ability to guide users in selecting, understanding, and interpreting fairness notions and metrics, we apply the Fairness Metrics Ontology as a tool to assist with the analysis of synthetic data fairness. It can analyze the results of a model generated on the synthetic data and recommend ML fairness metrics which should be evaluated, creating a reliable resource for synthetic fairness evaluation where the choice of metrics is now identified by the ontology rather than the individual researcher. Furthermore, the observed results can be compared with fairness of real datasets to highlight any new biases observed in synthetic data.

To illustrate this use case, we generated a Logistic Regression (LR) model on a subset of the MIMIC-3 [34] dataset based on a previously published study of synthetic data. The goal of the model is to predict 30-day patient mortality while achieving the best balanced accuracy scores on our validation data. The model is trained on 70% of the MIMIC-3 data, and then validated on 15% of the remaining data. The remaining 15% is used as the test dataset and we present the results for the same. The model’s results are passed to the Fairness Ontology, by representing those results in knowledge graph format as described in Section 4.2.3. We then query the ontology for fairness metrics which are used to evaluate the synthetic data—the query used is the top right one shown in Figure 5, so that we specifically use fairness metrics for notions which 1) are derived from false negative measurements, 2) can be used to address representation bias. Additionally, we 3) filter out results which are incompatible with the results of the MIMIC-3 synthetic dataset model—as “balance for positive class” requires a probability score not represented in the knowledge graph, this fairness notion is filtered out. With the remaining fairness notions, we use the corresponding fairness difference metric in order to generate fairness scores for the model. We repeat the experiment 10 times and average the results of fairness scores. The fairness scores are evaluated as combinations of subgroups of protected attributes: (a) age, (b) race and (c) gender.
In this case study of mortality prediction, the Fairness Ontology recommended “Equal Opportunity” and “Equalized Odds” as the two metrics to evaluate. For the current problem, the two fairness metrics are quite appropriate. They are based on the True Positive Rate (TPR) difference and the False Positive Rate (FPR) difference between the two subgroups of protected attribute. TPR identifies how correctly did the model predict the mortality of people who actually died. This is useful to ensure that appropriate care is provided to these individuals. FPR identifies how incorrectly did the model predict the mortality of people who actually did not die. This is important to identify so that we do not neglect people who are likely to survive. Thus, the ontology very aptly captured the two aspects of the fairness in the current classification problem and suggested the metrics. Additionally, as equalized odds is represented as a superclass of equal opportunity—as shown in the query explanation at the bottom of Figure 5—we decided to query the ontology an additional time, this time retrieving all fairness metrics which are sub-classes of equalized odds, and the ontology returned “Predictive Equality” as an answer. We did this in order to better understand the results of the first two metrics—as equal opportunity only relies on TPR, and equalized odds relies on the combination of TPR and FPR, we wanted to be able to observe any other fairness metrics which influence the end result of equalized odds—and the ontology was able to provide us with predictive equality, as the metric based on FPR alone. We present the results for Equal Opportunity Difference (TPR difference) and Predictive Equality (FPR difference) as they can simply be extended to Equalized Odds which is the maximum of the absolute values of Equal Opportunity Difference and Predictive Equality.

In the fairness context, the “privileged group” is considered the one with observed historical systematic advantage [4]. Similarly, the “unprivileged group” would represent those with historical systematic disadvantage. However, these vary based on the dataset and the problem. Females were the unprivileged group in the Adult income dataset while Males were privileged. This however was reversed for the COMPAS dataset where Males are considered as unprivileged [4]. As a result, for completeness in analysis on datasets with unknown historical biases, the analysis can be made robust by considering all possible unprivileged and privileged groups. Let the whole population be defined by \((x, s, y) \in \mathcal{D}_{\text{data}}\) where \(x\) represents the set of unprotected attributes, \(s\) represents the set of protected attributes and \(y\) represents the class label. We then identify two sub groups such that \(s = i\) becomes the unprivileged class for a random value \(i\) of the protected attribute \(s\) and \(s \neq i\) becomes the privileged class. For example, if unprivileged group represents the
population of black women, then privileged group would include all individuals who are not black women.

Figure 6 describes the results for the Equal Opportunity Difference and Predictive Equality for various protected attributes in the dataset using sunburst plots, similar to previous fairness studies [6, 29]. These insights are useful as taking the maximum of the absolute values of the two metrics results in Equalized Odds. The innermost ring describes the fairness of different age groups. The second ring in the plots stratifies the age groups along with race. For example, in the plots, the race value “Unknown” with its parent age group in the inner ring “<=45” represents the population of individuals who have age less than or equal to 45 and race Unknown. This becomes the unprivileged sub group while the rest of the population becomes the privileged group. This can further be extended to genders in the third ring. The sunburst plots were generated using Plotly [21] after the fairness metrics were identified by the ontology to visually describe the biases in the dataset, across combinations of protected attributes. While these plots were created after the metric identification, as a future direction, we plan to combine the visualization with the ontology to create an automated tool for fairness evaluation.

The fairness results are quite interesting. We note that at the high-level overview, the age group 81+ and its stratification by race and gender are the most favored by the model while all other age groups experience unfavorable predictions. Note that what is defined as a “favored group” changes depending on the fairness metric—according to Equal Opportunity Difference, a group is favored if it has a higher TPR than the other groups, while according to Predictive Equality Difference, a group is favored if its FPR is higher. Sunburst plots allow us to go deeper in our evaluation and thus, let’s also look at combination of protected attributes. While age group <=45 had unfavorable predictions overall, the Asians in this subgroup are being favored in terms of TPR. We already saw that individuals 81 and older are being favored towards. But the sub-group of these individuals who are males with their race unknown are actually experiencing unfavorable predictions by the model. This highlights that stratification by combination of protected attributes is essential and can reveal biases. Similar insights can also be drawn from the Predictive Equality results.

From using the ontology on synthetic data models, we can conclude that the ontology is not only able to represent the results of a model but also suggest ML fairness metrics which are useful for the problem at hand. This makes the use of the ontology as a potential solution for identifying fairness metrics for synthetic data. The robustness of the ontology can be highlighted using experiments on multiple datasets but this has been left as a future direction of this work.

6 CONCLUSION

We propose a Fairness Metrics Ontology, a knowledge resource for guiding users in selecting and understanding fairness metrics. We have built FMO to be usable in a variety of ways. Firstly, the ontology itself can be browsed through by a user, as we provide English-language definitions and usage examples for each fairness notion and metric that can help a user understand their purposes and distinctions. Secondly, using reasoning-based queries, a user can determine which fairness metrics meet certain criteria, such as whether a given fairness notion can be derived from a certain underlying metric or which combinations of metrics are equivalent to other metrics. Thirdly, using FMO’s ML concepts, a user can represent an ML system in an RDF knowledge graph, and use this information to infer which metrics can or cannot be computed based on available data.

In the future, we plan to continue to add to the variety of information stored within FMO, as well as demonstrate its utility with a specific use case. We also plan to build a fully-fledged system using the Whyis knowledge graph management and analysis framework [26] to demonstrate how FMO can be used to track and compute fairness information directly from model evaluation results in a provenance-sensitive way. By bringing FMO to the ML community, we provide a comprehensive knowledge resource for understanding, modeling, and reasoning about fairness in machine-learning models and datasets.

ACKNOWLEDGMENTS

This work is partially supported by IBM Research AI through the AI Horizons Network. We thank our colleague from IBM Research, Ching-Hua Chen, who greatly assisted the research and document preparation.

REFERENCES


[19] Steve Harris and Andy Seaborne (Eds.). 2013. A ONLINE RESOURCES


A ONLINE RESOURCES

The Fairness Metrics Ontology is available online at: <https://github.com/CognitiveHorizons/RPI-HEALS-fairness-metrics-ontology>.