The Lifecycle of "Facts": A Survey of Social Bias in Knowledge Graphs

Angelie Kraft[1] and **Ricardo Usbeck**[12]
[1]Department of Informatics, Universität Hamburg, Germany
[2]Hamburger Informatik Technologie-Center e.V. (HITeC), Germany
{angelie.kraft, ricardo.usbeck}@uni-hamburg.de

Abstract

Knowledge graphs are increasingly used in a plethora of downstream tasks or in the augmentation of statistical models to improve factuality. However, social biases are engraved in these representations and propagate downstream. We conducted a critical analysis of literature concerning biases at different steps of a knowledge graph lifecycle. We investigated factors introducing bias, as well as the biases that are rendered by knowledge graphs and their embedded versions afterward. Limitations of existing measurement and mitigation strategies are discussed and paths forward are proposed.

1 Introduction

Knowledge graphs (KGs) provide a structured and transparent form of information representation and lie at the core of popular Semantic Web technologies. They are utilized as a source of truth in a variety of downstream tasks (e.g., information extraction (Martínez-Rodríguez et al., 2020), link prediction (Getoor and Taskar, 2007; Ngomo et al., 2021), or question-answering (Höffner et al., 2017; Diefenbach et al., 2018; Chakraborty et al., 2021; Jiang and Usbeck, 2022)) and in hybrid AI systems (e.g., knowledge-augmented language models (Peters et al., 2019; Sun et al., 2020; Yu et al., 2022) or conversational AI (Gao et al., 2018; Gerritse et al., 2020)). In the latter, KGs are employed to enhance the factuality of statistical models (Athreya et al., 2018; Rony et al., 2022). In this overview article, we question the ethical integrity of these facts and investigate the lifecycle of KGs (Auer et al., 2012; Paulheim, 2017) with respect to bias influences.[1]

We claim that KGs manifest social biases and potentially propagate harmful prejudices. To utilize the full potential of KG technologies, such ethical risks must be targeted and avoided during development and application. Using an extensive literature analysis, this article provides a reflection on previous efforts and suggestions for future work.

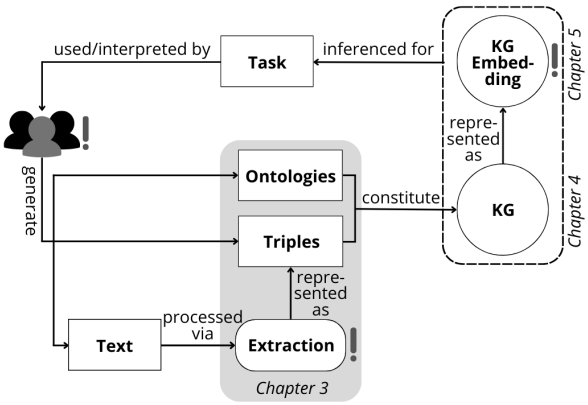

Figure 1: Overview of the knowledge graph lifecycle as discussed in this paper. Exclamation marks indicate factors that introduce or amplify bias. We examine bias-inducing factors of triple crowd-sourcing, hand-crafted ontologies, and automated information extraction (Chapter 3), as well as the resulting social biases in KGs (Chapter 4) and KG embeddings, including approaches for measurement and mitigation (Chapter 5).

We collected articles via Google Scholar[2] and filtered for titles including *knowledge graph/base/resource*, *ontologies*, *named entity recognition*, or *relation extraction*, paired with variants of *bias*, *debiasing*, *harms*, *ethical*, and *fairness*. We selected peer-reviewed publications (in journals, conference or workshop proceedings, and book chapters) from 2010 onward, related to social bias in the KG lifecycle. This resulted in a final count of 18 papers. Table 1 gives an overview of the reviewed works and Figure 1 illustrates the analyzed lifecycle stages.

[1]We focus on the KG lifecycle from a bias and fairness lens. For reference, the processes investigated in Section 3 correspond to the *authoring stage* in the taxonomy by Auer et al. (2012). The representation issues in KGs (Section 4) and KG embeddings (Sections 5 and 7) which affect downstream task bias relate to Auer et al.'s *classification stage*.

[2]A literature search on Science Direct, ACM Digital Library, and Springer did not provide additional results.

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 639–652
November 20–23, 2022.

2 Notes on Bias, Fairness, and Factuality

In the following, we clarify our operational definitions of the most relevant concepts in our analysis.

2.1 Bias

If we refer to a model or representation as *biased*, we — unless otherwise specified — mean that the model or representation is *socially biased*, i.e., biased towards certain social groups. This is usually indicated by a systematic and unfairly discriminating deviation in the way members of these groups are represented compared to others (Friedman and Nissenbaum, 1996) (also known as *algorithmic bias*). Such bias can stem from pre-existing societal inequalities and attitudes, such as prejudice and stereotypes, or arise on an algorithmic level, through design choices and formalization (Friedman and Nissenbaum, 1996). From a more impact-focused perspective, algorithmic bias can be described as "a skew that [causes] harm" (Kate Crawford, Keynote at NIPS2017). Such harm can manifest itself in unfair distribution of resources or derogatory misrepresentation of a disfavored group. We refer to *fairness* as the absence of bias.

2.2 Unwanted Biases and Harms

One can distinguish between *allocational* and *representational harms* (Barocas et al., as cited in, Blodgett et al., 2020), where the first refers to the unfair distribution of chances and resources and the second more broadly denotes types of insult or derogation, distorted representation, or lack of representation altogether. To quantify biases that lead to representational harm, analyses of more abstract constructs are required. Mehrabi et al. (2021a), for example, measure indicators of representational harm via *polarized perceptions*: a predominant association of groups with either negative or positive prejudice, denigration, or favoritism. Polarized perceptions are assumed to correspond to societal stereotypes. They can *overgeneralize* to all members of a social group (e.g., "*all* lawyers are dishonest"). It can be said that harm is to be prevented by avoiding or removing algorithmic bias. However, different views on the conditions for fairness can be found in the literature and, in consequence, different definitions of *unwanted* bias.

2.3 Factuality versus Fairness

We consider a KG factual if it is representative of the real world. For example, if it contains only male U.S. presidents, it truthfully represents the world as it is and has been. However, inference based on this snapshot would lead to the prediction that people of other genders cannot or will not become presidents. This would be false with respect to U.S. law and/or undermine the potential of non-male persons. Statistical inference over historical entities is one of the main usages of KGs. The factuality narrative, thus, risks consolidating and propagating pre-existing societal inequalities and works against matters of social fairness. Even if the data represented are not affected by sampling errors, they are restricted to describing *the world as it is* as opposed to *the world as it should be*. We strive for the latter kind of inference basis. Apart from that, in the following sections we will learn that popular KGs are indeed affected by sampling biases, which further amplify societal biases.

3 Entering the Lifecycle: Bias in Knowledge Graph Creation

We enter the lifecycle view (Figure 1) by investigating the processes underlying the creation of KGs. We focus on the human factors behind the authoring of *ontologies* and *triples* which constitute KGs. Furthermore, we address automated *information extraction*, i.e., the detection and extraction of entities and relations from text, since these approaches can be subject to algorithmic bias.

3.1 Triples: Crowd-Sourcing of Facts

Popular large-scale KGs, like Wikidata (Vrandecic and Krötzsch, 2014) and DBpedia (Auer et al., 2007) are the products of continuous crowd-sourcing efforts. Both of these examples are closely related to Wikipedia, where the top five languages (English, Cebuano, German, Swedish, and French) constitute 35% of all articles on this platform.[3] It can be said that Wikipedia is Euro-centric in tendency. Moreover, the majority of authors are white males.[4] As a result, the data transport a particular homogeneous set of interests and knowledge (Beytía et al., 2022; Wagner et al., 2015). This *sampling bias* affects the geospatial coverage of information (Janowicz et al., 2018) and leads to higher barriers for female personalities to receive

[3] https://en.wikipedia.org/wiki/List_of_Wikipedias

[4] https://en.wikipedia.org/wiki/Gender_bias_on_Wikipedia;https://en.wikipedia.org/wiki/Racial_bias_on_Wikipedia

a biographic entry (Beytía et al., 2022). In an experiment, Demartini (2019) asked crowd contributors to provide a factual answer to the (politically charged) question of whether or not Catalonia is a part of Spain. The diverging responses indicated that participants' beliefs of what counts as true differed largely. This is an example of bias that is beyond a subliminal psychological level. In this case, structural aspects like consumed media and social discourse play an important role. To counter this problem, Demartini (2019) suggests actively asking contributors for evidence supporting their statements, as well as keeping track of their demographic backgrounds. This makes underlying motivations and possible sources for bias traceable.

3.2 Ontologies: Manual Creation of Rules

Ontologies determine rules regarding allowed types of entities and relations or their usage. They are often hand-made and a source of bias (Janowicz et al., 2018) due to the influence of opinions, motivations, and personal choices (Keet, 2021): Factors like scientific opinions (e.g., historical ideas about race), socio-culture (e.g., how many people a person can be married to), or political and religious views (e.g., classifying a person of type X as a *terrorist* or a *protestor*) can proximately lead to an encoding of social bias. Also structural constraints like the ontologies' granularity levels can induce bias (Keet, 2021). Furthermore, issues can arise from the types of information used to characterize a person entity. Whether one attributes the person with their skin color or not could theoretically determine the emergence of racist bias in a downstream application (Paparidis and Kotis, 2021). Geller and Kollapally (2021) give a practical example for detection and alleviation of ontology bias in a real-world scenario. The authors discovered that ontological gaps in the medical context lead to an under-reporting of race-specific incidents. They were able to suggest countermeasures based on a structured analysis of real incidents and external terminological resources.

3.3 Extraction: Automated Extraction of Information

Natural language processing (NLP) methods can be used to recognize and extract entities (named entity recognition; NER) and their relations (relation extraction; RE), which are then represented as [head entity, relation, tail entity] tuples (or as [subject, predicate, object], respectively).

Mehrabi et al. (2020) showed that the NER system CoreNLP (Manning et al., 2014) exhibits binary gender bias. They used a number of template sentences, like "<Name> is going to school" or "<Name> is a person" using male and female names[5] from 139 years of census data. The model returned more erroneous tags for female names. Similarly, Mishra et al. (2020) created synthetic sentences from adjusted Winogender (Rudinger et al., 2018) templates with names associated with different ethnicities and genders. A range of different NER systems were evaluated (bidirectional LSTMs with Conditional Random Field (BiLSTM CRF) (Huang et al., 2015) on GloVe (Pennington et al., 2014), ConceptNet (Speer et al., 2017) and ELMo (Peters et al., 2017) embeddings, CoreNLP, and spaCy[6] NER models). Across models, non-white names yielded on average lower performance scores than white names. Generally, ELMo exhibited the least bias. Although ConceptNet is debiased for gender and ethnicity[7], it was found to produce strongly varied accuracy values.

Gaut et al. (2020) analyzed binary gender bias in a popular open-source neural relation extraction (NRE) model, OpenNRE (Han et al., 2019). For this purpose, the authors created a new dataset, named WikiGenderBias (sourced from Wikipedia and DBpedia). All sentences describe a gendered subject with one of four relations: *spouse*, *hypernym*, *birthData*, or *birthPlace* (DBpedia mostly uses occupation-related hypernyms). The most notable bias found was the spouse relation. It was more reliably predicted for male than female entities. This observation stands in contrast to the predominance of female instances with spouse relation in WikiGenderBias. The authors experimented with three different mitigation strategies: downsampling the training data to equalize the number of male and female instances, augmenting the data by artificially introducing new female instances, and finally word embedding debiasing (Bolukbasi et al., 2016). Only downsampling facilitated a reduction of bias that did not come at the cost of model performance.

Nowadays, contextualized transformer-based encoders are used in various NLP applications, includ-

[5]While most of the works presented here refer to gender as a binary concept, this does not agree with our understanding. We acknowledge that gender is continuous and technology must do this reality justice.

[6]https://spacy.io/

[7]https://blog.conceptnet.io/posts/2017/conceptnet-numberbatch-17-04-better-less-stereotyped-word-vectors/

ing NER and NRE. Several works have analyzed the various societal biases encoded in large-scale word embeddings (like word2vec (Mikolov et al., 2013; Bolukbasi et al., 2016) or BERT (Devlin et al., 2019; Kurita et al., 2019)) or language models (like GPT-2 (Radford et al., 2019; Kirk et al., 2021) and GPT-3 (Brown et al., 2020; Abid et al., 2021)). Thus, it is likely that these biases also affect the downstream tasks discussed here. Li et al. (2021) used two types of tasks to analyze bias in BERT-based RE on the newly created Wiki80 and TACRED (Zhang et al., 2017) benchmarks. For the first task, they masked only entity names with a special token (*masked-entity*; ME), whereas for the second task, only the entity names were given (*only-entity*; OE). The model maintained higher performances in the OE setting, indicating that the entity names were more informative of the predicted relation than the contextual information. This hints at what the authors call *semantic bias*.

A Note on Reporting Bias Generally, when extracting knowledge from text, one should be aware that the frequency with which facts are reported is not representative of their real-world prevalence. Humans tend to mention only events, outcomes, or properties that are out of their perceived ordinary (Gordon and Van Durme, 2013) (e.g., "a banana is yellow" is too trivial to be reported). This phenomenon is called *reporting bias* and likely stems from a need to be as informative and non-redundant as possible when sharing knowledge.

4 Bias in Knowledge Graphs

Next in our investigation of the lifecycle (Figure 1) comes the representation of entities and relations as a KG. In the following, we illustrate which social biases are manifested in KGs and how.

4.1 Descriptive Statistics

Janowicz et al. (2018) demonstrated that DBpedia, which is sourced from Wikipedia info boxes, mostly represents the western and industrialized world. Matching the coverage of location entries in the KG with population density all over the world showed that several countries and continents are underrepresented. A disproportionate 70% of the person entities in Wikidata are male (20% are female, less than 1% are neither male nor female, and for roughly 10% the gender is not indicated) (Beytía et al., 2022). Radstok et al. (2021) found that the most frequent occupation is *researcher* and Beytía et al. (2022) identified *arts*, *sports*, and *science and technology* as the most prominent occupation categories. In reality, only about 2% of people in the U.S. are researchers (Radstok et al., 2021). This gap is likely caused by reporting bias as discussed earlier (Section 3.3). Radstok et al. (2021), moreover, observed that mentions of ethnic group membership decreased and changed in focus between the 18th and 21st century. Greeks are the most frequently labeled ethnic group among historic entries (over 400 times) and African Americans among modern entries (only roughly 100 times).

4.2 Semantic Polarity

Mehrabi et al. (2021b) focused on biases in common sense KGs like ConceptNet (Speer et al., 2017) and GenericsKB (Bhakthavatsalam et al., 2020) (contains sentences) which are at risk of causing representational harms (see Section 2.2). They utilized *regard* (Sheng et al., 2019) and *sentiment* as intermediate bias proxies. Both concepts express the polarity of statements and can be measured via classifiers that predict a neutral, negative, or positive label (Sheng et al., 2019; Dhamala et al., 2021). Groups that are referred to in a mostly positive way are interpreted as favored and vice versa. Mehrabi et al. (2021b) applied this principle to natural language statements generated from ConceptNet triples. They found that subject and object entities relating to the professions *CEO*, *nurse*, and *physician* were more often favored while *performing artist*, *politician*, and *prisoner* were more often disfavored. Similarly, several Islam-related entities were on the negative end while *Christian* and *Hindu* were more ambiguously valuated. As for gender, no significant difference was found.

5 Bias in Knowledge Graph Embeddings

Vector representations of KGs are used in a range of downstream tasks or combined with other types of neural models (Nickel et al., 2016; Ristoski et al., 2019). They facilitate efficient aggregation of connectivity patterns and convey latent information.

Embeddings are created through statistical modeling and summarize distributional characteristics. So, if a KG like Wikidata contains mostly (if not only) male presidents, the relationship between the gender *male* and the profession *president* is assumed to manifest itself accordingly in the model. In fact, the papers summarized below provide evidence that the social biases of KGs are modeled or

further amplified by KG embeddings (KGEs). The following sections are organized by measurement strategy to give an overview of existing approaches and the information gained from them.

5.1 Stereotypical Analogies

The idea behind analogy tests is to see whether demographics are associated with attributes in stereotypical ways (e.g., "Man is to computer programmer as woman is to homemaker" (Bolukbasi et al., 2016)). In their in-depth analysis of a TransE-embedded Wikidata KG, Bourli and Pitoura (2020) investigated occupational analogies for binary gender seeds. TransE (Bordes et al., 2013) represents (h, r, t) (with head h, relation r, tail t) in a single space such that $h+r \approx t$. The authors identified the model's most likely instance of the claim "*a* is to *x* as *b* is to *y*" (with *(a,b)* being a set of demographics seeds and *(x,y)* a set of attributes) via a cosine score: $S_{(a,b)}(x, y) = cos(\vec{a} + \vec{r} - \vec{b}, \vec{x} + \vec{r} - \vec{y})$, where r is the relation *has_occupation*. In their study, the highest scoring analogy was "woman is to fashion model as man is to businessperson". This example appears rather stereotypical, but other highly ranked analogies less so, like "Japanese entertainer" versus "businessperson" (Bourli and Pitoura, 2020). A systematic evaluation of how stereotypical the results are is missing here. In comparison, the work that originally introduced analogy testing for word2vec (Bolukbasi et al., 2016) employed human annotators to rate stereotypical and gender-appropriate analogies (e.g., "sister" versus "brother").

5.2 Projection onto a Bias Subspace

Projection-based measurement of bias is another approach that was first proposed by Bolukbasi et al. (2016) for word embeddings, and was adapted for TransE by Bourli and Pitoura (2020). In a first step, a one-dimensional gender direction $\vec{d_g}$ is extracted. Then, a projection score metric S is computed to indicate gender bias — with projection π of an occupation vector $\vec{o}$ onto $\vec{d_g}$ and a set of occupations C: $S(C) = \frac{1}{|C|} \sum_{o \in C} ||\pi_{\vec{d_g}} \vec{o}||$. Occupations with higher scores are interpreted as more gender-biased and those with close-to-zero scores as neutral.

5.3 Update-Based Measurement

The *translational likelihood* (TL) metric was tailored for translation-based modeling approaches (Fisher et al., 2020b). To compute this metric, the embedding of a person entity is updated for one step towards one pole of a seed dimension. This update is done in the same way as the model was originally fit in. For example, if head entity *person x* is updated in the direction of *male* gender, the TL value is given by the difference between the likelihood of *person x* being a *doctor* after versus before the update. If the absolute value averaged across all human entities is high, this indicates a bias regarding the examined seed-attribute pair. Fisher et al. (2020b) argue that this measurement technique avoids model-specificity as it generalizes to any scoring function. However, Keidar et al. (2021) found that the TL metric does not compare well between different types of embeddings (details in Section 6). It should, thus, only be used for the comparison of biases within one kind of representation. Du et al. (2022) propose an approach comparable to Fisher et al. (2020b) to measure individual-level bias. Instead of updating towards a gender dimension, the authors suggest flipping the entity's gender and fully re-training the model afterward. The difference between pre- and post-update link prediction errors gives the bias metric. A validation of the approach was done on TransE for a Freebase subset (FB5M (Bordes et al., 2015)) (Du et al., 2022). The summed per-gender averages (group-level metric) were found to correlate with U.S. census gender distributions of occupations.

6 Downstream Task Bias: Link Prediction

Link prediction is a standard downstream task that targets the prediction of relations between entities in a given KG. Systematic deviations in the relations suggested for entities with different demographics indicate reproduced social bias.

For the measurement of fairness or bias in link prediction, Keidar et al. (2021) distinguish between *demographic parity* versus *predictive parity*. The assumption underlying demographic parity is that the equality between predictions for demographic counterfactuals (opposite demographics, for example, *female* versus *male* in binary understanding) is the ideal state (Dwork et al., 2012). That is, the probability of predicting a label should be the same for both groups. Predictive parity is given, on the other hand, if the probability of true positive predictions (*positive predictive value* or *precision*) is equal between groups (Chouldechova, 2017). Hence, this measure factors in the label distribution by demographic.

Table 1: Overview of reviewed works concerning the sources, measurement, and mitigation of bias in KGs/KGEs.

Bias Source		
Crowd-Sourcing		Beytía et al. (2022); Janowicz et al. (2018); Demartini (2019)
Ontologies		Janowicz et al. (2018); Keet (2021); Paparidis and Kotis (2021); Geller and Kollapally (2021)
Extraction		Mehrabi et al. (2020); Mishra et al. (2020); Gaut et al. (2020); Li et al. (2021)
Bias Measurement		
Representation	*Method*	
KG	Descriptive Statistics	Janowicz et al. (2018); Radstok et al. (2021); Beytía et al. (2022)
	Semantic Polarity	Mehrabi et al. (2021b)
KGE	Analogies	Bourli and Pitoura (2020)
	Projection	Bourli and Pitoura (2020)
	Update-Based	Fisher et al. (2020b); Keidar et al. (2021); Du et al. (2022)
	Link Prediction	Keidar et al. (2021); Arduini et al. (2020); Radstok et al. (2021); Du et al. (2022)
Bias Mitigation		
Representation	*Method*	
KGE	Data Balancing	Radstok et al. (2021); Du et al. (2022)
	Adversarial Learning	Fisher et al. (2020a); Arduini et al. (2020)
	Hard Debiasing	Bourli and Pitoura (2020)

With these metrics, Keidar et al. (2021) analyzed different embedding types, namely TransE, ComplEx, RotatE, and DistMult, each fit on the benchmark datasets FB15k-237 (Toutanova and Chen, 2015) and Wikidata5m (Wang et al., 2021). They averaged the scores across a large set of human-associated relations to detect automatically which relations are most biased. The results showed that *position played on a sports team* was most consistently gender-biased across embeddings. Arduini et al. (2020) analyzed link prediction parity regarding the relations *gender* and *occupation* to estimate debiasing effects on TransH (Wang et al., 2014) and TransD (Ji et al., 2015). The comparability between different forms of vector representations is a strength of downstream metrics. In contrast, measures like the analogy test or projection score (Bourli and Pitoura, 2020) are based on specific distance metrics and TL (Fisher et al., 2020b) was shown to lack transferability across representations (Keidar et al., 2021) (Section 5.3).

Du et al. (2022) interpret the correlation between gender and link prediction errors as an indicator of group bias. With this, they found, for example, that *engineer* and *nurse* are stereotypically biased in FB5M. However, the ground truth gender ratio was found not predictive of the bias metric (e.g., despite its higher male ratio, *animator* produced a stronger female bias value). For validation, it was shown that the predicted bias values correlate to the gender distributions of occupations according to U.S. census (again, on TransE). Furthermore, the authors investigated how much single triples contribute to group bias via an *influence function*. They found that gender bias is mostly driven by triples containing gendered entities and triples of low degree.

7 Breaking the Cycle? Bias Mitigation in Knowledge Graph Embeddings

A number of works have attempted to post-hoc mitigate biases in KGEs. Given that pre-existing biases are hard to eradicate from KGs, manipulating embedding procedures, may alleviate the issue at least on a representation level. In the following, we summarize respective approaches.

7.1 Data Balancing

Radstok et al. (2021) explored the effects of training an embedding model on a gender-balanced subset of Wikidata triples. First, the authors worked with the originally gender-imbalanced Wikidata12k (Leblay and Chekol, 2018; Dasgupta et al., 2018) and DBpedia15k (Sun et al., 2017) on which they fit a TransE and a DistMult model (Yang et al., 2015). They then added more female triples from the Wikidata/DBpedia graph to even out the binary gender distribution among the top-5 most common occupations. Through link prediction, they compared the number of male and female predictions with the ground truth frequencies. More female entities were predicted after the data balancing intervention. However, the absolute difference between the female ratios in the data and the predictions increased, causing the model to be less accurate and fair. Moreover, the authors note that this process is not scalable since for some domains there are no or only a limited amount of female entities (e.g., female U.S. presidents do not exist in Wikidata).

Du et al. (2022) experimented with adding and removing triples to gender-balance a Freebase subset (Bordes et al., 2015). For the first approach, the authors added synthetic triples (as opposed to real entities from another source as was done by Radstok et al. (2021)) for occupations with a higher male ratio. The resulting bias change was inconsis-

tent across occupations. This appears in line with the authors' finding that ground truth gender ratios are not perfectly predictive of downstream task bias (Section 6). For the second strategy, the triples that most strongly influenced an existing bias were determined and removed. This outperformed random triple removal.

7.2 Adversarial Learning

Adversarial learning for model fairness aims to prevent prediction of a specific personal attribute from a person's entity embedding. As an adversarial loss, Fisher et al. (2020a) used the KL-divergence between the link prediction score distribution and an idealized target distribution. For example, for an even target score distribution for a set of religions, the model is incentivized to give each of them equal probability. However, in their experiments, this treatment failed to remove the targeted bias fully. This is likely caused by related information encoded in the embedding that is able to inform the same bias.

Arduini et al. (2020) used a Filtering Adversarial Network (FAN) with a filter and a discriminator module. The filter intends to remove sensitive attribute information from the input, while the discriminator tries to predict the sensitive attribute from the output. Both modules were separately pre-trained (filter as an identity mapper of the embedding and discriminator as a gender predictor) and then jointly trained as adversaries. In their experiments, the gender classification accuracy for high- and low-degree entities was close to random for the filtered embeddings (TransH and TransD). For an additional occupation classifier, accuracy remained unaffected after treatment.

7.3 Hard Debiasing

Bourli and Pitoura (2020) propose applying the projection-based approach explained in Section 5.2 for the debiasing of TransE occupation embeddings. To achieve this, its linear projection onto the previously computed gender direction is subtracted from the occupation embedding. A variant of this technique ("soft" debiasing) aims to preserve some degree of gender information by applying a weight $0 < \lambda < 1$ to the projection value before subtraction. In the authors' experiments, the correlation between gender and occupation was effectively removed — as indicated by the projection measure (Bourli and Pitoura, 2020). However, the debiasing degree determined by λ was found to be in trade-off with model accuracy. This technique was closely adapted from Bolukbasi et al. (2016), regarding which Gonen and Goldberg (2019) criticize that gender bias is only reduced according to their specific measure and not the "complete manifestation of this bias".

8 Discussion

In this article, we cover a wide range of evidence for harmful biases at different stages during the lifecycle of "facts" as represented in KGs. Some of the most influential graphs misrepresent *the world as it is* due to sampling and algorithmic biases at the creation step. Pre-existing biases are exaggerated in these representations. Embedding models learn to encode the same or further amplified versions of these biases. Since the training of high-quality embeddings is costly, they are, in practice, pre-trained once and afterward reused and fine-tuned for different systems. These systems preserve the inherited biases over long periods, exacerbating the issue further. Our survey shows that KGs may qualify as resources for historic facts, but they do not qualify for inference regarding various human attributes. Future work on biases in KGs and KGEs should aim for improvement in the following areas:

Attribute and Seed Choices Bias metrics usually examine one or a few specific attributes (e.g., occupation) and their correlations with selected seed dimensions (e.g., gender). Occupation is by far the most researched attribute in the articles we found (Arduini et al., 2020; Radstok et al., 2021; Bourli and Pitoura, 2020; Fisher et al., 2020a,b). Only Keidar et al. (2021) propose to aggregate the correlations between a set of seed dimensions and all relations in a graph. All the works used binary gender as the seed dimension and some additionally addressed ethnicity, religion, and nationality (Fisher et al., 2020a,b; Mehrabi et al., 2021b).

Lack of Validation Most of the KGE bias metrics presented here are interpreted as valid if they detect unfairly discriminating association patterns that intuitively align with existing stereotypes. Besides that, several works investigate the comparability between different metrics. Although both of these practices deliver valuable information on validity, they largely ignore the societal context. Only Du et al. (2022) compared embedding-level bias metrics with census-aligned data to assess compatibility with real-world inequalities. We suggest that

future work consider a more comprehensive study of *construct validity* (Does the measurement instrument measure the construct in a meaningful and useful capacity?) (Jacobs and Wallach, 2021). One requirement is that the obtained measurements capture all relevant aspects of the construct the instrument claims to measure. That is, a gender bias measure must measure all relevant aspects of gender bias (Stanczak and Augenstein, 2021) (including, e.g., nonbinary gender and a distinction between benevolent and hostile forms of sexist stereotyping (Glick and Fiske, 1997)). Unless proven otherwise, we must be skeptical that this is achieved by existing approaches (Gonen and Goldberg, 2019). As a result of minimal validation, detailed interpretation guidelines are generally not provided. Therefore, the distinctions between strong and weak bias or weak bias and random variation are mostly vague.

(In-)Effectiveness of Mitigation Strategies Data balancing is the most intuitive approach to bias mitigation and was proven to be effective in the context of text processing (Meade et al., 2022). However, for KGEs, data balancing methods were found to inconsistently reduce bias (Section 7.1). Adversarial learning yielded promising outcomes in the study by Arduini et al. (2020). Their FAN approach does not rely on pre-specified attributes. This is in contrast to Fisher et al. (2020a), whose intervention was found to miss non-targeted, yet bias-related information. This problem relates to one of the main criticisms of hard and soft debiasing: instead of alleviating the problem, these techniques risk concealing the full extent of the bias (Gonen and Goldberg, 2019).

Reported Motivations Many, yet not all works in the field name potential social harms as a motivator for their research on social bias in KGs (Mehrabi et al., 2021b; Fisher et al., 2020a,b; Radstok et al., 2021). Only Mehrabi et al. (2021b) drew from established taxonomies and targeted biases associated with *representational harms* (Barocas et al., as cited in, Blodgett et al., 2020). Similarly, most works lack a clear working definition of social bias. For example, aspects of pre-existing societal biases captured in the data and biases arising through the algorithm (Friedman and Nissenbaum, 1996) are usually not disentangled. Only Bourli and Pitoura (2020) compared model bias to the original KG frequencies and showed that the statistical modeling caused an amplification.

9 Recommendations

To avoid harms caused by biases in KGs and their embeddings, we identify and recommend several actions for practitioners and researchers.

Transparency and Accountability KGs should by default be published with bias-sensitive documentation to facilitate transparency and accountability regarding potential risks. *Data Statements* (Bender and Friedman, 2018) report curation criteria, language variety, demographics of the data authors and annotators, relevant indicators of context, quality, and provenance. *Datasheets for Datasets* (Gebru et al., 2021) additionally state motivation, composition, preparation, distribution, and maintenance. The associated questionnaire can accompany the dataset creation process to avoid risks early on. Especially in the case of ongoing crowdsourcing efforts for encyclopedic KGs the demographic background of contributors should be reported (Demartini, 2019). Researchers using subsets of these KGs, should investigate respective data dumps for potential biases and report limitations transparently. Similarly, KG embedding models should be published with *Model Cards* (Mitchell et al., 2019) documenting intended use, underlying data, ethical considerations, and limitations. Stating the contact details for reporting problems and concerns establishes accountability (Mitchell et al., 2019; Gebru et al., 2021).

Improving Representativeness To tackle selection bias, data collection should aim to employ authors and annotators from diverse social groups and with varied cultural imprints. Annotations should be determined via aggregation (see Hovy and Prabhumoye, 2021). For open editable KGs, interventions like *edit-a-thons* are helpful to introduce more authors from underrepresented groups (Vetter et al., 2022) (e.g., the Art+Feminism campaign aims to fill the gender gap in Wikimedia knowledge bases[8]). In order for such interventions to take effect, research must update data bases and benchmarks frequently (see Koch et al., 2021). In addition, the timeliness of encyclopedic data is necessary to avoid perpetuating historic biases.

Tackling Algorithmic Bias Evaluation and prevention of harmful biases must become part of the development pipeline (Stanczak and Augenstein,

[8] https://outreachdashboard.wmflabs.org/campaigns/artfeminism_2022/overview

2021). Algorithmic biases are best evaluated with a combination of multiple quantitative (Section 5) and qualitative measures (Kraft et al., 2022; Dev et al., 2021), considering multiple demographic dimensions (beyond gender and occupation). Evaluating the content of attributions in light of social discourse and the intended use of a technology facilitates an assessment of potential harms (Selbst et al., 2019). Downstream task bias may exist independently from a measured embedding bias (Goldfarb-Tarrant et al., 2021), therefore a task- and context-oriented evaluation is preferred (Section 6). We have presented several bias-mitigating strategies for different KGEs, which might alleviate the issue in some cases (Section 7). However, more research is needed to establish more effective and robust mitigation methods, as well as metrics used to evaluate their impact (Gonen and Goldberg, 2019; Blodgett et al., 2020).

10 Related Work

Although a wide range of surveys investigates biases in NLP, none of them addresses KG-based methods, in particular. Blodgett et al. (2020) critically investigated the theoretical foundation of works analyzing bias in NLP. The authors claim that most works lack a clear taxonomy. We came to a similar conclusion with respect to evaluations of KGs and their embeddings. Sun et al. (2019) and Stanczak and Augenstein (2021) surveyed algorithmic measurement and mitigation strategies for gender bias in NLP. Sheng et al. (2021) summarized approaches for the measurement and mitigation of bias in generative language models. Some of the methods presented earlier are derived from works discussed in these surveys and adapted to the constraints of KG embeddings (e.g., Bourli and Pitoura (2020) adapted hard debiasing (Bolukbasi et al., 2016)). Criticisms point to the monolingual focus on the English language, the predominant assumption of a gender binary, and a lack of interdisciplinary collaboration.

Shah et al. (2020) identified four sources of predictive biases: *label bias* (label distributions are imbalanced and erroneous regarding certain demographics), *selection bias* (the data sample is not representative of the real world distribution), *semantic bias/input representation bias* (e.g., feature creation with biased embeddings), and *overamplification* through the predictive model (slight differences between human attributes are overemphasized by the model). All of these factors are reflected in the lifecycle as discussed in this article. To counter the risks, Shah et al. (2020) suggest employing multiple annotators and methods of aggregation (see also Hovy and Prabhumoye, 2021), re-stratification, re-weighting, or data augmentation, debiasing of models, and, finally, standardized data and model documentation.

11 Conclusion and Paths Forward

Our survey shows that biases affect KGs at different stages of their lifecycle. Social biases enter KGs in various ways at the creation step (e.g., through crowd-sourcing of triples and ontologies) and manifest in popular graphs, like DBpedia (Beytía et al., 2022) or ConceptNet (Mehrabi et al., 2021b). Embedding models can capture exaggerated versions of these biases (Bourli and Pitoura, 2020), which finally propagate downstream (Keidar et al., 2021). We acknowledge that KGs have enormous potential for a variety of knowledge-driven downstream applications (Martínez-Rodríguez et al., 2020; Ngomo et al., 2021; Jiang and Usbeck, 2022) and improvements in the truthfulness of statistical models (Athreya et al., 2018; Rony et al., 2022). Yet, although KGs are factual about historic instances, they also perpetuate historically emerging social inequalities. Thus, ethical implications must be considered when developing or reusing these technologies.

We showed that most embedding-based measurement approaches for bias are still restricted to a limited number of demographic seeds and attributes. Furthermore, their alignment with social bias as a construct is not sufficiently validated. Some debiasing strategies appear effective within rather narrow definitions of bias. More in-depth scrutiny is required for a broader understanding of bias. Future work should be grounded in an investigation of concepts like gender or ethnic bias and strive for more comprehensive operationalizations and validation studies. Finally, the motivations and conceptualizations should be communicated clearly.

Acknowledgments

We acknowledge the financial support from the Federal Ministry for Economic Affairs and Energy of Germany in the project CoyPu (project number 01MK21007[G]) and the German Research Foundation in the project NFDI4DS (project number 460234259).

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Large language models associate Muslims with violence. *Nature Machine Intelligence*, 3(6):461–463.

Mario Arduini, Lorenzo Noci, Federico Pirovano, Ce Zhang, Yash Raj Shrestha, and Bibek Paudel. 2020. Adversarial learning for debiasing knowledge graph embeddings. In *Proceedings of the 16th International Workshop on Mining and Learning with Graphs (MLG)*.

Ram G. Athreya, Axel-Cyrille Ngonga Ngomo, and Ricardo Usbeck. 2018. Enhancing community interactions with data-driven chatbots–the DBpedia chatbot. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 143–146, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer.

Sören Auer, Lorenz Bühmann, Christian Dirschl, Orri Erling, Michael Hausenblas, Robert Isele, Jens Lehmann, Michael Martin, Pablo N. Mendes, Bert Van Nuffelen, Claus Stadler, Sebastian Tramp, and Hugh Williams. 2012. Managing the life-cycle of linked data with the LOD2 stack. In *The Semantic Web - ISWC 2012 - 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part II*, volume 7650 of *Lecture Notes in Computer Science*, pages 1–16. Springer.

Emily M. Bender and Batya 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.

Pablo Beytía, Pushkal Agarwal, Miriam Redi, and Vivek K. Singh. 2022. Visual gender biases in Wikipedia: A systematic evaluation across the ten most spoken languages. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 43–54.

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. 2020. GenericsKB: A knowledge base of generic statements. *CoRR*, abs/2005.00660.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 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*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *CoRR*, abs/1506.02075.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Styliani Bourli and Evaggelia Pitoura. 2020. Bias in knowledge graph embeddings. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 6–10.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Nilesh Chakraborty, Denis Lukovnikov, Gaurav Maheshwari, Priyansh Trivedi, Jens Lehmann, and Asja Fischer. 2021. Introduction to neural network-based question answering over knowledge graphs. *WIREs Data Mining and Knowledge Discovery*, 11(3):e1389.

Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163.

Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha Talukdar. 2018. HyTE: Hyperplane-based temporally aware knowledge graph embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2001–2011, Brussels, Belgium. Association for Computational Linguistics.

Gianluca Demartini. 2019. Implicit bias in crowdsourced knowledge graphs. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 624–630, San Francisco, USA. Association for Computing Machinery.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Nanyun Peng, and Kai-Wei Chang. 2021. What do bias measures measure? *CoRR*, abs/2108.03362.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 862–872, Online. Association for Computing Machinery.

Dennis Diefenbach, Vanessa López, Kamal Deep Singh, and Pierre Maret. 2018. Core techniques of question answering systems over knowledge bases: a survey. *Knowledge and Information Systems*, 55(3):529–569.

Yupei Du, Qi Zheng, Yuanbin Wu, Man Lan, Yan Yang, and Meirong Ma. 2022. Understanding gender bias in knowledge base embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1381–1395, Dublin, Ireland. Association for Computational Linguistics.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, Cambridge, Massachusetts. Association for Computing Machinery.

Joseph Fisher, Arpit Mittal, Dave Palfrey, and Christos Christodoulopoulos. 2020a. Debiasing knowledge graph embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7332–7345, Online. Association for Computational Linguistics.

Joseph Fisher, Dave Palfrey, Christos Christodoulopoulos, and Arpit Mittal. 2020b. Measuring social bias in knowledge graph embeddings. In *Proceedings of the AKBC 2020 Workshop on Bias in Automatic Knowledge Graph Construction*.

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems*, 14(3):330–347.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '18, page 1371–1374, Ann Arbor, MI, USA. Association for Computing Machinery.

Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2020. Towards understanding gender bias in relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2943–2953, Online. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

James Geller and Navya Martin Kollapally. 2021. Detecting, reporting and alleviating racial biases in standardized medical terminologies and ontologies. In *Proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1–5.

Emma J. Gerritse, Faegheh Hasibi, and Arjen P. de Vries. 2020. Bias in conversational search: The double-edged sword of the personalized knowledge graph. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, ICTIR '20, page 133–136, Online. Association for Computing Machinery.

Lise Getoor and Ben Taskar. 2007. *Introduction to Statistical Relational Learning*. The MIT Press.

Peter Glick and Susan T Fiske. 1997. Hostile and benevolent sexism: Measuring ambivalent sexist attitudes toward women. *Psychology of Women Quarterly*, 21(1):119–135.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 60–63, Florence, Italy. Association for Computational Linguistics.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, page 25–30, San Francisco, California, USA. Association for Computing Machinery.

Xu Han, Tianyu Gao, Yuan Yao, Deming Ye, Zhiyuan Liu, and Maosong Sun. 2019. OpenNRE: An open and extensible toolkit for neural relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 169–174, Hong Kong, China. Association for Computational Linguistics.

Konrad Höffner, Sebastian Walter, Edgard Marx, Ricardo Usbeck, Jens Lehmann, and Axel-Cyrille Ngonga Ngomo. 2017. Survey on challenges of question answering in the semantic web. *Semantic Web*, 8(6):895–920.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8).

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 375–385, Online. Association for Computing Machinery.

Krzysztof Janowicz, Bo Yan, Blake Regalia, Rui Zhu, and Gengchen Mai. 2018. Debiasing knowledge graphs: Why female presidents are not like female popes. In *Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks co-located with 17th International Semantic Web Conference (ISWC 2018)*.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 687–696, Beijing, China. Association for Computational Linguistics.

Longquan Jiang and Ricardo Usbeck. 2022. Knowledge graph question answering datasets and their generalizability: Are they enough for future research? In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3209–3218, Madrid, Spain. Association for Computing Machinery.

C. Maria Keet. 2021. An exploration into cognitive bias in ontologies. In *Joint Ontology Workshops 2021 Episode VII: The Bolzano Summer of Knowledge, JOWO 2021*.

Daphna Keidar, Mian Zhong, Ce Zhang, Yash Raj Shrestha, and Bibek Paudel. 2021. Towards automatic bias detection in knowledge graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3804–3811, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 2611–2624. Curran Associates, Inc.

Bernard Koch, Emily Denton, Alex Hanna, and Jacob G Foster. 2021. Reduced, reused and recycled: The life of a dataset in machine learning research. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Angelie Kraft, Hans-Peter Zorn, Pascal Fecht, Judith Simon, Chris Biemann, and Ricardo Usbeck. 2022. Measuring gender bias in german language generation. In *INFORMATIK 2022*, pages 1257–1274. Gesellschaft für Informatik, Bonn.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Julien Leblay and Melisachew Wudage Chekol. 2018. Deriving validity time in knowledge graph. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1771–1776, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Luoqiu Li, Xiang Chen, Hongbin Ye, Zhen Bi, Shumin Deng, Ningyu Zhang, and Huajun Chen. 2021. On robustness and bias analysis of BERT-based relation extraction. In *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction*, pages 43–59, Singapore. Springer Singapore.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

José-Lázaro Martínez-Rodríguez, Aidan Hogan, and Ivan López-Arévalo. 2020. Information extraction meets the semantic web: A survey. *Semantic Web*, 11(2):255–335.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2020. Man is to person as woman is to location: Measuring gender bias in named entity recognition. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, HT '20, page 231–232, Online. Association for Computing Machinery.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021a. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6).

Ninareh Mehrabi, Pei Zhou, Fred Morstatter, Jay Pujara, Xiang Ren, and Aram Galstyan. 2021b. Lawyers are dishonest? Quantifying representational harms in commonsense knowledge resources. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5033, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–12.

Shubhanshu Mishra, Sijun He, and Luca Belli. 2020. Assessing demographic bias in named entity recognition. In *Proceedings of the AKBC 2020 Workshop on Bias in Automatic Knowledge Graph Construction*.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, Atlanta, GA, USA. Association for Computing Machinery.

Axel-Cyrille Ngonga Ngomo, Mohamed Ahmed Sherif, Kleanthi Georgala, Mofeed Mohamed Hassan, Kevin Dreßler, Klaus Lyko, Daniel Obraczka, and Tommaso Soru. 2021. LIMES: A framework for link discovery on the semantic web. *Künstliche Intelligenz*, 35(3):413–423.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2016. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33.

Evangelos Paparidis and Konstantinos Kotis. 2021. Towards engineering fair ontologies: Unbiasing a surveillance ontology. In *2021 IEEE International Conference on Progress in Informatics and Computing (PIC)*, pages 226–231.

Heiko Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 8(3):489–508.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765, Vancouver, Canada. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Wessel Radstok, Melisachew Wudage Chekol, and Mirko T. Schäfer. 2021. Are knowledge graph embedding models biased, or is it the data that they are trained on? In *Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021)*.

Petar Ristoski, Jessica Rosati, Tommaso Di Noia, Renato De Leone, and Heiko Paulheim. 2019. RDF2Vec: RDF graph embeddings and their applications. *Semantic Web*, 10(4):721–752.

Md. Rashad Al Hasan Rony, Ricardo Usbeck, and Jens Lehmann. 2022. DialoKG: Knowledge-structure aware task-oriented dialogue generation. *CoRR*, abs/2204.09149.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 59–68, Atlanta, GA, USA. Association for Computing Machinery.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4444–4451.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *CoRR*, abs/2112.14168.

Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. 2020. CoLAKE: Contextualized language and knowledge embedding. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3660–3670, Online. International Committee on Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Zequn Sun, Wei Hu, and Chengkai Li. 2017. Cross-lingual entity alignment via joint attribute-preserving embedding. In *The Semantic Web – ISWC 2017*, pages 628–644, Cham. Springer International Publishing.

Kristina Toutanova and Danqi Chen. 2015. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 57–66, Beijing, China. Association for Computational Linguistics.

Matthew A. Vetter, Krista Speicher Sarraf, and Elin Woods. 2022. Assessing the art+ feminism edit-a-thon for Wikipedia literacy, learning outcomes, and critical thinking. *Interactive Learning Environments*, 30(6):1155–1167.

Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Claudia Wagner, David García, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 454–463, Oxford, UK. AAAI Press.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1112–1119, Québec City, Québec, Canada. AAAI Press.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022. A survey of knowledge-enhanced text generation. *ACM Computing Surveys*. Just Accepted.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.