A Systematic Survey of Semantic Web Technologies for Bias in Artificial Intelligence Solutions

Paula Reyero [a,*], Enrico Daga [a], Harith Alani [a] and Miriam Fernandez [a]
[a] *Knowledge Media Institute, The Open University, United Kingdom*
E-mails: paula.reyero-lobo@open.ac.uk, harith.alani@open.ac.uk, miriam.fernandez@open.ac.uk

Abstract. Bias in artificial intelligence (AI) is a critical and timely issue due to its sociological, economic and legal impact, as decisions made for humans by algorithms could lead to unfair treatment of certain individuals or groups of individuals. Multiple surveys have emerged to give a multidisciplinary overview of bias [1–3] or to review bias in applied areas such as social sciences [4–6], business research [7], criminal justice [8], or data mining [9–14]. Due to the capability of Semantic Web (SW) technologies to fulfil data validity gaps in many AI areas [15], we revise the extent to which they can contribute to bringing solutions to this problem. To the best of our knowledge, there exists no previous work to bring together bias and semantics, so we review their intersectionality following a systematic approach [16]. Consequently, we provide in-depth analysis and categorisation of different types and sources of bias addressed with semantic approaches and discuss their advantages to improve frequent limitations in AI systems. We find works in the areas of information retrieval, recommendation systems, machine and deep learning, and natural language processing, and argue through multiple use cases that semantics can help especially dealing with technical, sociological, and psychological challenges.

Keywords: Bias in AI, Conceptual Semantics, Semantic Web technologies, Algorithmic fairness

1. Introduction

AI systems are widely used in society as they provide considerable benefits to both individuals and business, but their decisions have occasionally shown to reproduce or even amplify human biases [17]. There is increasingly more awareness of the problems of bias and discrimination in numerous AI applications, and the direction is consequently shifting towards pursuing, not only accurate, but ethical AI systems [18].

One of the main advantages of these systems over human intelligence is their ability to process huge amounts of data. It is clear how crucial data is for building intelligent systems that act accordingly with the *intention* they were designed for. It has been claimed that an algorithm is only as good as the data it works with [17]. Dealing with data validity gaps such as data being incomplete, not representative, and erroneous is one of the main challenges when building AI systems. However, other factors coming from the humans that use these systems and their interaction can propagate bias [6], so minimising these factors is a complex and fundamental task in AI system design.

The vast amount of information available in the SW has a huge potential to fill in existing gaps of AI systems, leveraging the structured formalisation of knowledge that is machine-understandable to build more realistic and fairer models. There exist examples in different areas, including machine learning and data mining, natural language processing, or social networks and media representation, in which the Semantic Web, Linked Data and Web of Data have made a significant contribution [15].

Many scholars have raised attention to the lack of consistency between the motivation and the technological solutions proposed to address bias [14]. There is

*Corresponding author. E-mail: paula.reyero-lobo@open.ac.uk.

a need to conceptualise bias in terms of what system behaviours are considered harmful, in which ways, to whom, and why. The aim of this survey is to provide a review of the contribution of Semantic Web technologies in the context of bias following an in-depth conceptualisation that accounts for these dimensions. For this purpose, we refer to the existing literature to provide a categorisation at different stages of the AI processing pipeline building upon existing frameworks. This work is relevant to the SW and to the broader AI scientific communities, as bias is gaining attention in different areas including computer science, social sciences, philosophy and law [18].

We follow a systematic approach [16] to review the related literature of semantic methods for developing solutions to different biases in AI. Our main contribution in this survey is: (i) to give a comprehensive and critical overview of state-of-the-art techniques based on semantics used to identify, capture, or mitigate bias, (ii) to highlight the advantages of using different semantic approaches considering the bias type, origin, and impact, and (iii) to identify the main opportunities and pitfalls (limitations of bias evaluation methods and bias within semantic resources) in the intersectional literature of semantics and bias.

The rest of the paper is organised as follows. In Section 2, we define the SW technologies considered in this survey (conceptual semantics). In Section 3, we describe the methodology followed in the systematic literature review. Section 4 describes the semantic and bias dimensions found in this survey's categorisation, followed by the analysis of the surveyed papers in Section 5. Finally, Section 6 is a discussion of the main findings and Section 7 are some conclusion remarks.

2. Background of conceptual semantics

This survey follows the definition of semantic resources as *formal, structured, and standardised data structures, which make explicit the meaning of the information by extracting the concepts and explicit relations between them* [19]. In increasing order of complexity, these include taxonomies, thesauri, ontologies, knowledge bases, knowledge graphs, and the linked data.

There exist different approaches to extract the meaning of text by machines and another one commonly used is distributional semantics [20]. However, we note that distributional word embeddings, lexicon-based algorithms and similar methods that only rely on word patterns are not the focus of this work.

Taxonomies. Taxonomies are hierarchical or faceted structures that group words according to their meaning to provide a classification of concepts. For example, SKOS [19] is a common data model for sharing and linking knowledge organisation systems in the SW.

Thesauri. Thesaurus add standard structured relationships and other properties to each concept, such as related and alternative terms. An example of this kind is WordNet [21], a commonly known lexical database which presents the sense of words in a relational structure.

Ontologies. Following Gruber's definition [22], an ontology is a formal, explicit specification of a shared conceptualisation that is characterised by high semantic expressiveness required for increased complexity. A general objective is to integrate data from different sources, as ontologies must follow the World Wide Web Consortium (W3C) standard format and are linked to the SW.

Knowledge graphs. A knowledge graph (KG) is based on a formal knowledge representation of the data as a graph, that is, a network of nodes and links representing concepts, classes, properties, relationships and entity descriptions. For example, ConceptNet [23] is an example of a KG of 1.6 million assertions of commonsense knowledge (e.g., "cooking food can be fun", represented in the graph as <cook food> <capableOf> <be fun>).

Linked Data. Finally, we refer to Linked Open Data (LOD) as structured data available on the Web in a standard format, to be reachable and manageable by SW tools, e.g., by semantic queries [24].

Many data objects can fit in more than one of the given definitions. However, in this article we refer to various data objects according to the definition given in the corresponding source paper.

3. Survey methodology

To provide a thorough literature review, we followed the guidelines of the systematic mapping study research method [16] to conduct research that is inclusive of the key primary studies in the domain.

This section presents the methodology to plan (Section 3.1), execute (Section 3.2) and synthesise (Section 3.3) the significant works.

3.1. Planning search strategy

Specify the research questions. The motivation of this study is bias and conceptual semantics.

Regarding semantics, we investigate methods, evaluation frameworks and metrics to investigate the utility of semantics to address bias problems (Section 4.1).

Regarding bias, we identify which type of biases can appear in AI systems (Section 4.2.1). Secondly, we discuss the most common sources of bias addressed with semantics (Section 4.2.2). Finally, we explore how bias manifest to identify key challenges in AI (Section 4.2.3).

Search string and database sources The collection of relevant studies is based on an extensive keyword-based querying of the two main elements of the survey (Table 1) in two popular scholarly databases: Elsevier Scopus and ISI Web of Knowledge (WoS). We complete our search with Microsoft Academic Search and Google Scholar for a snowballing process [25].

Search string
TITLE-ABS-KEY(‘bias*’ OR ’debias*’)
AND
TITLE-ABS-KEY(’knowledge graph*’ OR ’knowledge base*’
OR ’ontology’ OR ’ontologies’ OR ’ontological representation’
OR ’ontological knowledge’ OR ’thesaurus’ OR ’thesauri’
OR ’conceptual semantic*’)

Table 1
List of keywords include in the string search query. TITLE-ABS-KEY refer to the title, abstract and keywords of the paper, respectively.

3.2. Search execution

We collect significant works according to specific inclusion criteria and a filtering process (Figure 1).

Inclusion criteria (IC) definition

- IC1: Papers written in English.
- IC2: Studies published in relevant journals between 2010 and 2020.
- IC3: Only papers subjected to peer review, which include published journal papers, as part of conference proceedings or workshop, and book chapters.

Filtering selection

We remove duplicates from the two databases and we filter papers in four following steps.

Source-based filter Primary studies relevant are selected from the Scopus publication sources of Computer Science, Mathematics, Engineering, Business, Decision Science and Social Sciences, and completed with WoS.

Metadata-based filter Papers are selected based on title, abstract, publication venue and publication year.

We exclude not primary studies (e.g., project proposals and literature reviews) and papers published in more than one venue. Only the latest version available, the one with most complete results, or most relevant publication venue is included.

Content-based filter We screened papers based on the introduction, conclusion, or full-text, especially in unclear studies.

The focus of this survey is to investigate SW technologies used in solutions for bias coming from an AI system, not from the users or communities. Therefore, we discard papers that lack an AI system or the use of SW technologies. For example, some works lack evidence of improving bias in the AI system (e.g., [26] did not perform any relevant experiment or vision on how to address the problem of cognitive bias in recommendation systems) or do not use semantics in their solution (e.g., [27] use knowledge graph embeddings but bias is addressed using disjoint test classes), or use knowledge bases which are not conceptual [28].

Study-selection from snowballing We include additional studies from paper citations, when the filtered papers are read in more detail.

3.3. Synthesis of the results

The last step is the categorisation of filtered papers in a data extraction form according to the analysed features defined in Section 4.

4. Dimensions of analysis

In this section, we define bias and the conceptualisation used to organise the primary research works that have emerged from the literature.

Section 4.1 presents different categories of semantic approaches (dimensions of conceptual semantic tasks). Section 4.2 defines different categories of bias according to its type (Section 4.2.1), origin (Section 4.2.2), and impact (Section 4.2.3).

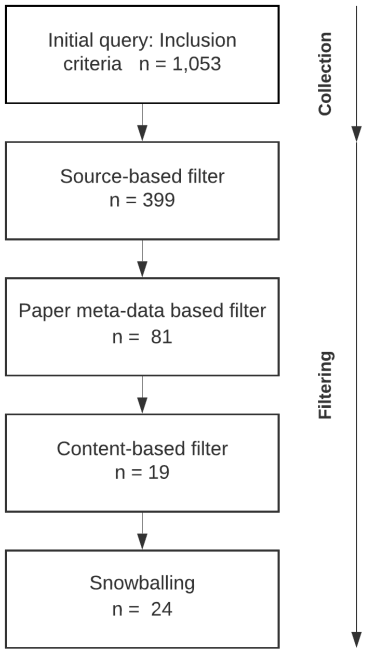

Fig. 1. Diagram showing the paper filtering detailed in Section 3.2.

4.1. Dimensions of conceptual semantic tasks

In this section, we present the three main group of works according to their use of semantics to address bias.

Identifying bias Semantics can be used to discover bias. As an example, [29] use the representation of user-item interactions in a graph to discover disparities in the recommendation quality between groups that have a different activity, e.g., due to being economically disadvantaged.

Modeling bias Semantics can be used to model bias as part of the AI system. For example, bias can reduce the performance of a system due to, e.g., limited coverage and diversity in the training data of how sentiment manifest across languages [30]. Semantics can be used to structure and capture bias to avoid misrepresentation of minority groups in the downstream applications.

Mitigating bias Finally, semantics can be used to reduce the impact of bias in AI systems. There are three main types of bias mitigation approaches [1, 4]: those focusing on changing the training data [31–38], the learning algorithm during the model generation [39–42], or the model outcomes based on a holdout set which was not involved during the training phase [29].

4.2. Bias in AI categorisation

There are many definitions of bias in AI systems but this survey follows the same definition as [4] of bias due to the heterogeneities in data, some of which we will mention below. *Data is generated by subgroups of people with their own characteristics and behaviors. These heterogeneities can bias the data. A model learned from biased data may lead to unfair and inaccurate predictions.*

In this context, we pose three questions. First, *how we can define the types of bias addressed using semantics* (Section 4.2.1). Second, *where these biases originate in the AI workflow* (Section 4.2.2). Third, *what are the challenges and how does bias impact the AI systems* (Section 4.2.3).

4.2.1. Types of bias addressed with semantics

To understand the bias types, we look at the three points of view used by R.Baeza-Yates to address bias and define their nature [6].

Statistical bias. From a statistical point of view, statistical bias is a systematic deviation of the, possibly unknown, real distribution of the variables represented in the data.

Cognitive bias. From a psychological perspective, cognitive bias appears due to the way humans make decisions in a subjective manner.

Cultural bias. From a sociological point of view, cultural bias is existing in the inclinations and beliefs in society that can reflect in the data.

4.2.2. Location of bias in the AI workflow

To understand the source of each bias, we follow Oltenau et al.'s [5] comprehensive framework of limitations and pitfalls when using data for decision-making.

Bias at source. The first critical point is at the data origin or source, as any bias existing at the input of an AI system will appear at least in the same way at the output. This is commonly referred as the "garbage in - garbage out" principle.

In the surveyed papers, this can be due to functional and external biases, as defined in Section 5.2.

Bias at collection. Next, we come across issues when collecting the data, which can happen due to inaccuracies either at the sampling, querying, or filtering processes.

Data pre-processing. Following this is the data preparation, which include errors during cleaning, annotation, or aggregation pre-processing techniques.

Data analysis. Finally, we consider issues when analysing the data, for example, due to concerns arising from the choice of a certain methodology, such as, using data as a source of hypotheses rather than a tool to test them, tailoring research based on data availability, or testing multiple hypothesis until finding a positive result.

4.2.3. Bias impact

The impact of algorithmic bias in AI systems is fundamental to quantify its scale and severity. We find examples of the main challenges identified in [5] and new challenges in some of the papers.

Population bias. Bias as the over or under representation of certain demographic groups or user characteristics.

Behavioural bias. Bias in differences on how users seek and assimilate information.

Content bias. Bias as lexical, syntactic, semantic and structural differences in the content produced by the users.

Temporal bias. Bias as variations on population behaviors over time, for example, influenced by seasonality or trends.

Two other categories are outside the scope of Olteanu et al.'s [5] framework:

Data quality. Bias impacting as data sparsity and noise.

Model overfitting. Bias that results in algorithms fit to a specific set of data points with low predictive power.

5. Analysis of results

The analysis of results is delivered in two ways. First, Section 5.1 analyses the surveyed papers according to the bias type and origin to provide an understanding of how similar bias problems affect to different AI systems.

Then, Section 5.2 discusses in more detail the SW technologies used to identify, capture or mitigate bias according to its implication in the system to provide specific methodology examples in the context of frequent limitations of AI systems that could help extrapolate them to similar problems in future research.

5.1. Bias and AI systems

We first define the different AI systems covered by in this survey, to then analyse them according to their bias type (Table 2) and origin (Table 3). A cross-table with all the dimensions of bias and semantics discussed in Section 5 is shown in Table 5.

Table 2
Paper classification according to their bias type, as defined in Section 4.2.1.

Type of bias	*#Papers*	*Reference*
Statistical bias	10	[29, 31–33, 37, 38, 40, 41, 43, 44]
Cultural bias	6	[30, 35, 36, 45–47]
Cognitive bias	8	[34, 39, 42, 48–52]

Table 3
Paper classification according to their origin of bias, as defined in Section 4.2.2.

Origin of bias	*due to*	*Reference*
Bias at source	External bias	[30, 36, 45–47, 49]
	Functional bias	[29, 31, 39, 43, 48]
Bias at collection	Sampling	[35]
	Querying	[33, 38, 40, 44]
Data pre-processing	Annotation	[32, 34, 37, 42, 52]
	Aggregation	[41]
Data analysis	Inference and prediction	[50, 51]

The following categories group the different AI systems covered in this survey: information retrieval (IR), recommender systems (RS), machine learning (ML), deep learning (DL), natural language processing (NLP), and intelligence activity.

IR refers to the group of studies that focus on methods to manage, analyze and retrieve information (e.g., images [33, 38], music [35], web documents [39, 45, 48, 49], or content in social media [40]) based on a prompt made by the user, i.e., a user's query. In other words, these systems are used to fetch information that is relevant to what the user is searching for.

RS are aimed to discover new information to the users. Based on collecting explicit or implicit information about them, a model is build to select which items will be more likely preferred. Recommendations can be made on property similarity with past items (content-based filtering) or based on their similarity with other users' feedback that is collected from historical user-item interactions (collaborative filtering) [31], or the combination of both or other approaches to address specific issues (hybrid RS). One example of hybrid

RS that appears in the surveyed papers uses a dialog system with the content-based recommender to decide based on explicit users' preferences, as the dialog can help with the "cold start" problem of having little information about new users [44]. Another example is knowledge-based recommendation, in which a knowledge base is used to represent user-item interactions and derive similarity metrics from the graph [29, 43]. This is similar to a KG completion problem, where new recommendations would be the prediction links in a completion task. Knowledge graph embeddings are useful for recommendation because they not only provide the vector representations of the user-item entities and relations, but can also give additional information such as their properties and entity types, as the structure and semantics of the KG is preserved.

ML groups works that use techniques based on predictive algorithms that aim to train models to learn from data and make predictions. It can be done in a supervised manner (classification tasks to, for example, predict the sense of a word [52]), or unsupervised (clustering to discover similar groups of, for instance, action concepts [47]).

DL is used to refer to a specific subgroup ML algorithms based on neural models, also known as "black-box" models (e.g., to make automatic annotation of images [32], or in text classification tasks [46]).

NLP is used to comprise the broad area that aims to enable computers to understand natural language data (from the extraction of relevant information from text data [37, 42], to more specific tasks, such as opinion mining and sentiment analysis tasks [30, 34, 41], or detecting hate speech [36]).

Intelligence activity. Finally, we refer as intelligence activity to group two groups that develop an AI system to support the search, interpretation, selection and display of information to draw conclusions from masses of data [50, 51].

Then, we introduce different AI system problems and how semantics have helped on each case, presenting the works by similarity of statistical (5.1.1), cultural (5.1.2), and cognitive (5.1.3) bias type and source.

5.1.1. Statistical bias

Statistical bias is defined as a systematic deviation of the, possibly unknown, real distribution of the variables represented in the data, which results from an inaccurate estimation or sampling process.

Statistical bias was found to emerge:

1. At source due to a *functional bias* (Section 5.1.1.1): restrictions given by the design of each AI system, e.g., the introduction of new features affects the use of previous ones and shapes the user behaviour [29, 31, 43].
2. At collection during *querying* (Section 5.1.1.2): due to possible data loss or bias in the resulting dataset, e.g., caused by the lack of expressivity or an incorrect query formulation [33, 38, 40, 44].
3. At the data pre-processing during *annotation* (Section 5.1.1.3): manual or semi-automatic annotations can exacerbate existing biases or introduce new biases, e.g., noisy labels due to poorly designed annotation guidelines [32, 37].
4. At the data pre-processing during *aggregation* (Section 5.1.1.4): resulting from the process of assembling to structure, organise, represent or transform data, for example, to increase the potential for inferring new facts [41].

5.1.1.1. Bias at source (functional bias) Bias at source due to a functional bias is addressed in 3 different RS [29, 31, 43].

Collaborative RS. System in [31] is a matrix factorization algorithm, and semantics are used to deal with the *small fraction of negative samples in the data*, as most interactions in collaborative filtering are implicit positive feedback (e.g., clicks, purchases). Items related to the ones from the positive interactions are assumed to be more likely to be known to the user and, therefore, to be true negatives. Using a KG could provide high-quality negative samples, i.e., informative (their introduction has a significant change in the model parameters), and factual (they are true negatives, as the user has known them before, but did not choose them).

Knowledge-based Collaborative RS. Semantics are used to reduce the bias and performance disparity in the system in [29] due to dominance of the data from the most *active over the less active users* (e.g., economically disadvantaged users), which causes less visibility of their historic user-item interaction data. A KG is used to model explicitly the recommendations in the form of reasoning paths and to impose fairness across users using the user-item path distributions to quantify richness (number of graph patterns of each user) and evenness (relative importance of each pattern across users).

Hybrid RS. A KG is used to model user-item interactions in [43] to reduce bias against *less popular items* due to the predominant interaction of users with only the few most popular items. The methodology is based on hybrid property-specific subgraphs,

that is, vector representations of the user and item entities considering one property at a time to model feedback accounting for only meaningful interaction similarities between their semantic properties. Considering only property-specific vector representations allowed to model subproperty specific interactions (e.g., movies being related in terms of starring actors even if not having the same subject), which could improve overall recommendation specially in data with low popularity bias in terms of serendipity (precision of recommended items after discarding the obvious ones) and novelty (precision of items that are unknown, as they are part of the long-tail of the catalog).

5.1.1.2. Bias at collection (querying) Bias at collection due to querying is addressed in a novel end-to-end framework consisting on a hybrid RS [44], and 2 different IR systems for image retrieval [33, 38] and information retrieval in social media [40].

Hybrid RS. The system in [44] is based on a content-based RS linked to a dialog system which uses a KG to deal with bias due to the *limited representation of the user's preferences* to only the items mentioned in the dialog. Their framework propagates entities to a KG to compute similarity between user with items not yet mentioned, to enrich the user's preference representation and generate more consistent responses in the dialog system.

Image Retrieval. A KG is used in a sentence-based image retrieval in [38] (pre-trained model of visual word detectors trained in a large-scale annotated images) to reduce bias due to the *restricted vocabulary of the visual word detectors* to only the captions in the training data. The relations from the KG with other concepts helps to trigger related visual detectors to return more relevant queries. On the other hand, general knowledge semantic resources are used in [33] to deal with bias due to *complex natural language queries* in a keyword-based image retrieval system. The set of concept names and properties of the image caption and query are extracted from the graph to raise accuracy of queries that include broad concepts.

Information retrieval. Similarly, two domain-specific ontologies are used in [40] to represent the data more explicitly to reduce *parsing mistakes* when dealing with understanding of knowledge across different domains. A semantic indexing categorises the information according to relevant domain categories to produce more relevant results.

5.1.1.3. Data pre-processing (annotation) Data pre-processing bias due to annotation is addressed in DL neural models used for automatic image captioning [32], and natural language understanding for robots [37].

Image Captioning. Semantics are used in [32] to enrich the word context vector of the detected objects in an automatic encoder (Recurrent-Convolutional Neural Network, R-CNN) - decoder (Long Short-Term Memory, LSTM) image captioning framework to boost the process of visual attention of *out-of-training words*. KG information is useful to describe implicit intentions with new entities (e.g., a "woman standing with her luggage" next to a sign, we can speculate she can be waiting for the bus) that can generate more meaningful automatic annotations.

Natural Language Understanding. A common-sense KB is used in [37] to reveal and reduce inconsistencies in an LSTM model used as a semantic parser for natural language understanding for robots due to the *small, domain-specific training corpus*. Shifts of the model towards undesired behaviours (i.e., giving attention to the incorrect words of the sentence) could be reduced to some extent using examples of sentences that are different but involve the same action from the external resource.

5.1.1.4. Data pre-processing (aggregation) Data pre-processing bias due to aggregation is addressed in NLP for a new sentiment value propagation method [41].

Sentiment Value Propagation. Bias of the system in [41] is due to the *imbalance between positive and negative seeds of the training data* that are used to infer sentiment values to other concepts, and gradually leads new values to shift towards the average value. A bias correction step is proposed to align sentiment values to the mean and standard deviation that can be computed from an external source with all the concepts in a manually annotated sentiment dictionary (Affective Norms for English Words) to prevent new concepts to shift this value towards the dominant polarity.

5.1.2. Cultural bias

Cultural bias is already existing in inclinations to our shared personal beliefs or issues in society that can sneak into the data that is being used and learned by the AI system. As it is the reflection of historical or social inequalities in AI, it has also been commonly referred as historical bias.

Cultural bias was found to emerge:

1. At source due to an *external bias* (Section 5.1.2.1): influence of factors outside the AI sys-

tem such as socioeconomic status, education, privacy concerns, language, personality and culture influences can affect the reliability and/or representativeness of the data, e.g., the social context is reflected in the dataset and limits the generalizability of the conclusions that can be drawn from it [30, 36, 45–47].
2. At collection during *sampling* (Section 5.1.2.2): data sampled is not representative of the whole population, e.g., sample size is too small and does not follow the distribution of the real population [35].

5.1.2.1. Bias at source (external bias) Bias at source due to an external bias is addressed in IR for information retrieval from web text snippets and quotations from online news sources [45]. In a ML clustering analysis to provide a novel action conceptualisation [47]. Finally, in DL text classification algorithms for hate speech detection (CNN) [36] and 4 general binary classification problems (CNN and recurrent bidirectional LSTM) [46], and an image sentiment prediction model (CNN) [30].

Web Information Retrieval. Semantics are used in [45] to identify media political bias, i.e., *different presentation of information* due to the reporter's own opinion and perceptions. The system allows to structure this information so news media outlets reporting biased stories can be identified using visualization techniques (e.g., heatmaps over polarities of opinions extracted from each news media on the same topic).

Clustering. A multilingual ontology is used in the clustering in [47] to capture concepts of the action domain in a common space that is *inclusive to their interpretation in different languages*. As there is a language-specific understanding of which verbs trigger a certain action, an ontology of action videos is used to identify groups of verbs by action type (by "typological closeness") that are language independent and thus constitute an inter-linguistic classification action domain.

Text Classification. Semantics are used in [36] to reduce bias due to an *over-generalised belief about a particular group of people* that leads to an imbalance in the training data that cause the model to shift towards incorrect predictions for these groups. External semantic information is used to generalise text referring to disadvantaged groups to counteract for their higher error rate. Besides, semantics are used to describe the domain of influential features in different text classification tasks in [46] to capture *biased model preferences imposed by the training data* (e.g, a painter not being likely to be in an exposition due to its nationality) in the form of predictive rules. A rule-based ML method over the KG can discover predictions made by the model on the grounds of protected input attributes due to learned shortcut correlations in the training data.

Image Sentiment Analysis. An ontology is build in [30] to capture language-biased sentiment adjectives due to the different *cultural interpretations of sentiment and emotion in images*. The ontology is used to achieve an dataset for visual sentiment analysis with a wider coverage and diversity of visual affect across 12 different languages to avoid biased downstream tasks to one predominant culture interpretation.

5.1.2.2. Bias at collection (sampling) Bias at collection due to sampling is addressed in an IR methodology to integrate cultural context in a music information retrieval system [35].

Music Information Retrieval. Information from the linked open data is proposed in [35] to reduce bias towards mainly market-driven popular music due to most data in commercial music platforms with *lack of diversity and culture-agnostic*. A multimodal knowledge base is proposed to enrich music audios with knowledge extracted from semantic web technologies to better contextualise them and reveal non-trivial, deeper relations between music entities that can improve similarity measures for music discovery and recommendation.

5.1.3. Cognitive bias

Cognitive bias is defined as a systematic error due to the way humans process and interpret information. The effect of this subjectivity impacts their behavior, the way they make decisions and their judgements.

Cognitive bias was found to emerge:

1. At source due to an *external bias* (Section 5.1.3.1) [49].
2. At source due to a *functional bias* (Section 5.1.3.2) in [39, 48].
3. At the data pre-processing during *annotation* (Section 5.1.3.3): [34, 42, 52].
4. At the data analysis at inference and prediction (Section 5.1.3.4): errors when drawing conclusions beyond the dataset under analysis [50, 51].

5.1.3.1. Bias at source (external bias) Bias at source due to an external bias is addressed in an IR new method to perform sentiment analysis on web search queries [49].

Information Retrieval. A thesaurus, SentiWordNet, is used in [49] to predict sentiment of user's search queries to identify bias about polarised and opinionated topics. Identifying this type of biased content improved quality of query recommendation in the long-tail by suggesting less popular queries with aligned sentiment to improve results relevance, and queries with opposite sentiment to improve diversity of opinions.

5.1.3.2. Bias at source (functional bias) Bias at source due to a functional bias is addressed in IR to investigate ranking and sorting bias in search engine functionalities [39, 48].

Web Information Retrieval. A Knowledge Graph Box is incorporated in the search engine's interface [39] to counter for the human's heuristic subjective processing of information. Two features are integrated into the search interface, a knowledge box with comprehensible factual information and a warning message, to improve user's knowledge and attitude towards a specific controversial topic (e.g., vaccination). The user's exposed to factual information were significantly more knowledgeable, less skeptical of vaccination and more critical of information quality than the ones without any KG box information after a web search simulation. Similarly, a KG as a visualisation interface was compared to a general hierarchical tree in [48] to explore and understand the influence of the presentation of search results. Their qualitative research reveals that the KG interface supports exploratory search in viewing source documents fewer times and faster without reducing quality and user satisfaction of the gathered information.

5.1.3.3. Bias at collection (annotation) Bias at collection due to annotation is addressed in a ML word sense disambiguation (WSD) task in [52], and in NLP sentiment analysis for opinion mining in [34] and information extraction [42].

Word Disambiguation. Bias of subjective manual annotations is investigated in [52]. They compare the precision of two lexicographers to a supervised word sense disambiguation algorithm, using the scenario in which context (i.e., the set of neighboring words that provide domain information) is the gold standard. WordNet and Corpus-dependent parameters provide context in the supervised approach. Their study shows that context is paramount for this task, specially for humans, as the absence of it tends to shift towards the most frequent sense of a word and leaves the annotator in an indecisive state.

Sentiment Analysis. To deal with bias of annotations in subjective tasks, a thesaurus is used in [34] to provide an additional score in sentiment classification of reviews that can be prone to mixed opinions and the use of complex language. Difficult writing patterns of the reviewers and their use of subjective language can bias the classification results, so the semantic resource can alleviate this problem relying of more objective information.

Information Extraction. An ontology is build in [42] to deal with subjective interpretations of the mathematical models used to explain weather conditions (e.g., maps. graphs, or textual information). They develop a proprietary ontological model to compare worded forecasts (subjective to mood, fatigue, and humor) with analytical information to assess how truthful the forecasts are. The system is able to detect important events in the text and describe the semantics and linguistic information about different atmospheric variables, to compare the numerical values of their properties with the observation data.

5.1.3.4. Data analysis (inference and prediction) Data analysis bias due to inference and prediction is addressed in intelligence analysis of the Intelligence Cycle [51] (planning and direction, collection, processing and exploitation, analysis and production, dissemination and integration, and evaluation and feedback), and Intelligence Analysis [50] (discovery of evidence, hypothesis, and arguments in the scientific method framework) due to the impartial judgment of human simplified information processing, as defined by Heuer [53].

Intelligence Cycle. *CBOntology* is developed in [51] to support assessment and reduce bias in intelligence work. The principle of this application ontology is based on using the principles and rules of consistent and predictable cognitive patterns to render them explicitly and support users in intelligence activity tasks. To this aim, they extract expert knowledge from a taxonomy of several thousand intelligence activities and capture more than 400 classes of cognitive biases based on string, semantic, logical, and topological matching similarity of existing ontologies.

Intelligence Analysis. *TIACRITIS* is a domain KB developed in [50] with a different approach that uses semantic technologies to represent all the reasoning steps, evidence, probabilistic assessments and assumptions in a collaborative effort of several analysts teams to recognise well-known analysts' biases and advice the user to counter for them.

5.2. Bias impact and use of semantics

Section 5.2 aims to provide a higher-level analysis, by discussing in more detail the semantic methodologies to address the bias problems introduced in Section 5.1 considering frequent bias implications in AI systems (Table 4). We only include in Table 4 the studies that use an external semantic resource, not those one building one for their particular needs [29, 30, 39, 48, 50, 51].

We select the most appropriate papers from each bias category. First, [36, 45–47] as bias due to an external factor and [31, 43] as functional bias, as they provide semantic methods that outperform alternative approaches. Second, [35] as sampling and [38, 44] at querying, as their methods are tested across a wider variety of samples. Third, [32, 37] as bias in annotation, as they are the least specific to one problem of the system and therefore can be more representative.

Conceptual semantics dimension	*Resource*	*Reference*
Identifying bias	FrameNet	[37]
	YAGO	[45]
Capturing bias	LOD	[35]
	IMAGACT	[47]
	Wikidata	[46]
	DBpedia	[44]
Mitigating bias	FrameNet	[37]
	WordNet	[36]
	ConceptNet	[32, 38]
	Freebase	[31]
	DBpedia	[43]

Table 4
Paper classification (n = 11) that use an external semantic resource to address bias according to their dimension of conceptual semantics, as defined in Section 4.1.

Population bias One of the most common implications of bias is the over- or under-representation of certain demographics or other user characteristics in the dataset used with respect to a target population. Some common issues happen when using data in which user demographics are not equally represented, due to the type of platform different users are more prone to use, or the way mechanisms within the same platform (e.g., the use of hashtags in Twitter) vary across groups. The use of unreliable proxy populations can compromise the whole system.

We find semantic approaches to identify [29], capture [35, 46], and mitigate [36] this bias.

Capturing population bias in [46] is based on association rule-mining over Wikidata entities that correspond to the influential tokens of the input data. Bias is captured in the predicted rules over concepts and properties of the KG, and appear in the form of counter-intuitive predictions, or predictions that were based on sensitive data. A second example is Koduri et al. [35], where cultural context of music recordings is captured providing contextual features extracted from the Linked Open Data (LOD) through Open Information Extraction (Open IE) from unstructured data at web-scale. Using context-based features in combination with the content-based features (extracted from the audio recordings) can be used to model culturally-relevant similarity measures, such as biographical information or social connections of two artists with similar singing patterns, and create what they called a culturally sound navigation space that is more relevant to the cultural background of the music.

Identifying population bias in [29] is based on deriving fairness metrics from the KG to impose diversity in recommendation, as there are underrepresented users due to activity disparities. Fairness constrains based on user-item interaction paths (path score and diversity score) can be used as a model regularisation parameter, to mitigate less quality and explanation diversity of users with less historical data.

Mitigating population bias in [36] is based a preprocessing technique of the training data. WordNet is used for data correction based on the replacement of words that refer to disadvantaged communities with their higher-level hierarchical words (i.e, hypernyms) to balance their distribution in the training data. The abstraction of knowledge can reduce the disproportion of false positives in the communities represented by these words that is caused by the huge amount of hateful content that has been generated against these communities.

Content bias We refer to lexical, syntactic, semantic and structural differences in the content produced by the users, which depends on factors such as the use of language being different across countries, contextual factors influencing how users talk, the difference in content between "expert" and regular users, or the propensity of certain populations to talk about certain topics.

We find semantic approaches to capture bias due to the use of language across countries [30, 47].

Capturing content bias in [30] relies on an ontology to capture differences in how visual affect is ex-

pressed and perceived across languages. They proof that there are indeed distinctions by comparing performance of fine-tuned language-specific models in a task to predict sentiment in an image dataset that has been collected using ANPs of another language. Namely, they can use this multilingual unified ontology to evaluate to which extend the visual sentiment of a given language can be predicted by sentiment models trained in other languages. Another example is an ontology of video actions (IMAGACT) used as a video-based disambiguation framework for a clustering algorithm in [47]. IMAGACT is used to derive a similarity matrix to categorise action concepts which are not language-specific, i.e., biased to a particular language. This interlinguistic clustering domain is possible through the annotation of the ontology video prototypes with verbs in ten different languages, which can be used to perform clustering of action concepts using their multilingual lexical features.

Behavioural bias Behaviour disparities can appear in a wide range of user actions, including how they connect and interact with each other, how they seek information, and how they create content. How users communicate with each other is influenced, e.g., by their shared relationship, how they find and interact with content depends in their interest, expertise, and information needs; and how they create content is defined by their self-selection bias towards what, when and how they choose to create it.

We find two semantic approaches to mitigate bias due to the way users seek for information, and how they process it [39, 48].

Mitigating behavioural bias in [39] is approached changing the search environment with a manipulated version of Google's knowledge graph box. A box containing a summary of the related topic extracted from Wikidata was made visible to participants to mitigate their deficient processing of health information found in Google. Furthermore, the effect of the search interface in how users find and interact with content is explored in [48]. A post-test evaluation is conducted to show how the presentation of results using a KG (constructed using an Open IE system) or a hierarchical tree affect information seeking. Their study reveals that user's leaning towards a particular spatial representation is a affected by their exploration needs, i.e., a KG interface is more useful when users are searching for specific, and not new, information.

Temporal bias Variations on population behaviors change over time due to, e.g., non-stationary patterns of participation on certain topics that can be triggered by current trends, seasonality, or other engineered efforts such as marketing campaigns.

We find a semantic approach to identify bias due the different perspectives over time of news reporters on the same event [45].

Identifying temporal bias in [45] uses semantics to acquire candidate opinions, in particular, to extract the opinion holder from the collected texts. For this purpose, they used YAGO to perform neural entity recognition and map each holder to a canonical name. Holders, topics, and opinions are stored in triples that can be used to compare and identify biased news reports.

Data quality bias Two main issues regarding data are sparsity and noise. We find examples of bias that results as sparse dataset, when few over all possible elements are captured in it. This results in a dataset that is easy to analyse on the "head" (frequent elements), but not on the "tail". On the other hand, noise is due to incomplete, corrupted, or datasets with erroneous content.

We find semantic approaches to mitigate bias that exhibits in data sparsity, as less popular items are more difficult to mine [43], and mitigate bias manifested as missing data, as negative feedback is not collected to be used by the RS [31].

Mitigating sparsity bias in [43] collects feedback separately for each property of the user/item DBpedia's entity to improve recommendation, as the vector representation of the property-specific subgraph embedding used to compute the similarity metrics between entities is taking into account the meaning of each property. Thus, the recommendation accuracy of less popular items improves.

Mitigating noise bias in [31] builds a negative sampler based on reinforcement learning over Freebase co-trained with the RS to infer a negative sample for each user-item interaction data. This demonstrated to improve the top-K recommendation and preference ranking tasks metrics of seven baseline methods, which were also using KGs, but only to leverage positive signals.

Model overfitting Bias can manifest in model overfitting, due to algorithms being fit to a specific set of data points that have, as a result, very low predictive power, i.e., they make substantial errors when predicting out-of-training data points.

We find that semantics have been used to capture [44], identify [37], and mitigate [32, 37, 38] bias.

Capturing model overfitting in [44] is based on a pre-procesing approach to capture users' interests with the propagation of items of the training data to DBPedia to expand their feature vector representation. A relational graph convolutional network (R-GCN) is used to encode structural and relational information of the neighboring nodes in the KG of the mentioned items that are already part of the training data, to enable the RS to consider the importance of items and non-item entities in the recommendation.

Identifying model overfitting in [37] is based on annotating data with FrameNet, to check inconsistency of the black-box model. This is done by detecting misalignment between the words of a sentence with a higher value in the model's attention layer, and the lexical unit for that sentence (that corresponds to the action triggered from that sentence in FrameNet). Based on an error analysis, they reveal latent patterns that the model follow that are completely unrelated to the linguistic theory, as the model gives higher attention to recurrent words that are not discriminatory in the respective sentence.

Mitigating model overfitting in [37] is done using a statistical correction strategy based on data augmentation. Examples from FrameNet, used as gold standard, are introduced in the training data to make the model more consistent with the Frame Semantics theory and help it to generalise beyond the training corpus. Differently, [32, 38] follow probability-based approaches to estimate the likelihood of out-of-training words ("undetectable words") to be on an image based on their semantic relations with other entities extracted from ConceptNet (CN). In [38], this problem is assessed in an image retrieval task, by extending the metric used in the object detector's aggregation function to account for words that are directly connected on the graph. In [32], a similar approach is assessed in an automatic captioning framework. Semantically related words of the objects detected in the image are injected into the output of the caption generator (LSTM layer), to augment the probability of some latent meaningful words that can allow for better generalisation.

6. Discussion

We use the proposed bias conceptualisation in a wide range of AI systems in Section 5 to bring to a discussion how successful the use of semantics to address bias in AI has been in the last decade (Section 6.1), how these methods relate to the literature of bias and fairness (Section 6.2), and the current limitations of existing approaches (Section 6.3).

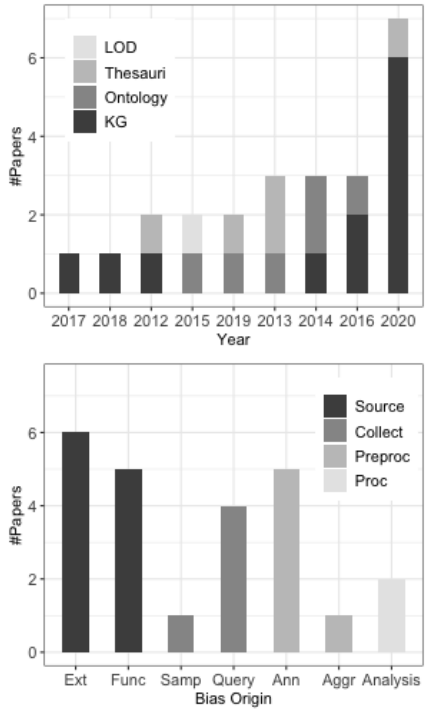

Fig. 2. Frequency distribution of studies in the time under scope (top). Frequency distribution according to the bias source as analysed in Section 5.2 (bottom). Abbreviations: LOD (Linked Open Data), KG (Knowledge Graph), Source (Bias at source), Collect (Bias at collection), Preproc (Data pre-processing), Proc (Data analysis).

6.1. *How beneficial are semantic resources to address bias in AI?*

In general, from the tables extracted from the literature we have the following findings:

- The work done to capture and mitigate bias using semantics has been more prominent than to identify them.
- As seen in Fig. 2, semantic resources have been used to address bias at different stages of the AI pipeline, especially to deal with bias at the data source.
- KGs have notably been used in the past year. ConceptNet [32, 33, 38, 41], DBpedia [43, 44], Wikidata [39, 46], Freebase [31], and YAGO [45] are the most common KGs.

Table 5
Classification of papers according to the bias type and origin and the semantic goal and resource. Bias type: statistical (ST), cultural (CU), and cognitive (CO). Semantics: identify (ID), mitigate (MI), capture (CA). Semantic resources: linked open data (LOD), knowledge graph (KG). AI systems: RS, recommender system; IR, information retrieval; NLP, natural language processing; ML, machine learning; DL, deep learning; IA, intelligence activity.

Bias type	*Bias source*	*Semantics*	*AI*	*Semantic resource*	*Ref.*
ST	Functional	ID	RS	-	[29]
		MI		KG	[31, 43]
	Querying	CA		Thesauri, KG	[44]
		MI	IR	KG	[33, 38]
				-	[40]
	Annotation	ID/MI	DL	Thesauri	[37]
		MI		KG	[32]
	Aggregation	MI	NLP	KG	[41]
CU	External	ID	IR	KG	[45]
		CA	ML	Ontology	[47]
			DL		[30]
				KG	[46]
		MI		Thesauri	[36]
	Sampling	CA	IR	LOD	[35]
CO	External	ID	IR	Thesauri	[49]
	Functional	MI		KG	[39]
				-	[48]
	Annotation	ID	ML	Thesauri	[52]
			NLP	-	[42]
		MI		Thesauri	[34]
	Analysis	ID	IA	-	[50]
		CA		-	[51]

On this basis, we argue the advantages of semantics in the following paragraphs.

Dealing with technical challenges. Semantics help with sampling bias in the training data. We have seen that KGs are particularly useful in **RS** to deal with imbalance in the sample data, due to differences in users' activity (disparities in the user data), item popularity (disparity in the item interactions), or type of feedback (of positive or negative samples). RS used similarity measures between entity [29] and property [43] paths to improve accuracy given imbalance of user and item data, respectively, and using neighbor entities as negative samples to deal with this missing data [31].

Furthermore, **neural models** have also benefited from semantics to deal with unintended bias of small and domain-specific samples. To expand knowledge from ground-truth annotations in the training data, related entities in the KG relevant for the task (automatic image annotation [32], neural propagation for recommendation [44], or image retrieval [38]) can be injected into the neural feature vector representations [32, 44] or as features to compute the score of the neural model [38]. Another approach is to increase the training data size using examples of a semantic resource as ground-truth annotations [37].

Dealing with sociological challenges. Semantics help with representation bias in the training data. Even if the sampling process is correct to fulfil the technical requirements of the AI system, data can indeed be not representative of some minority populations that are not as well represented as the general population.

- Ontologies can provide a common space for representation that is inclusive to different cultures, if using language as a proxy [30, 47].
- KGs can be populated with the training data to capture data disparities that reflect social issues, due to historical disadvantages [46], or bias due to the media sources [45].
- Semantic information is a mean to enrich multimedia data to capture its cultural context [35].

– Semantics can help correcting misrepresentations in data. For example, to correct for generalised beliefs and stereotypes in society, higher-level semantic information can be used to provide fairer distributions of disadvantaged communities and balance their data used in the AI task [36].

Dealing with psychological challenges. Finally, semantics can help with bias in the interpretation of the data shown to the user, especially when searching for information. KGs have been an advantageous search interface to counteract human's heuristic processing of information and mitigate impaired judgment and decision-making when drawing conclusions from huge amounts of data [39, 48].

In conclusion, we can say from a general point of view that semantics are useful in *capturing* bias.

– They can provide a richer representation of data, through which bias can be captured in the form of spurious correlations and relations extracted from patterns on the graph that would otherwise not be seen [46].
– Semantics can be used to extract contextual knowledge from the instances in the data, or additional properties and relations with instances outside the scope of the collected data [35, 43, 44, 46].
– The structure of the knowledge representation in ontologies can be useful to model bias due to different language interpretations [30, 47].

Semantics are also useful to *mitigate* bias:

– Ranking and sorting biases can be reduced by providing a structured representation of search results [39, 48].
– Semantics can reduce detrimental effects of data imbalances by adding related features [32, 38] or samples [31, 37], or correcting the features included in the dataset to account for bias [36].

6.2. How can these methods be framed in the literature of methods to address bias?

To discuss with more detail the strengths of semantic approaches, we group the works according to common methods used in the bias literature.

Imposing fairness metrics. Fairness metrics are one of the most prominent and established approaches in the current state of algorithmic fairness, to avoid biases and discrimination arising from the data or the algorithms in use. These methods rely on the use of measures that evaluate the outcome of the system in relation to some sensitive or protected attributes that should not impact the decision. However, to date there is not a universal understanding on how the notion of fairness should be defined. We see this in [4] and many other works, as there is a long list of fairness metrics and each one should be used depending on the application.

In the scope of this survey, we found an example of a fairness-aware algorithm in [29]. They proposed a post-processing approach relying on two fairness metrics to avoid individual and group discrimination of less active and disadvantages users. Their metrics were extracted from the path in KG used to represent the user-item interactions, and imposed in the ranking method used to calculate the relevance score of the user-item pairs. Group fairness is measured with the diversity score of the user-item path distributions, measured with the Simpson's Index of Diversity. Individual fairness relies on inequality dispersion in terms of consumption or income distribution, measured with the *Gini* coefficient.

However, the majority of semantic approaches found in this survey address bias from a pre-processing approach, that is, by looking into the data independently from the model. We introduce them in the subsequent paragraphs.

Dataset augmentation. Dataset augmentation is based in statistical correction through the increase of samples in the dataset. Methods in this category are based in pre-processing the data by adding new sampled data to deal with, for example, imbalanced distributions, or small, domain-specific datasets. General preprocessing approaches aim to modify the training set before learning to avoid discrimination by suppression (removing the attributes most correlated to the sensitive attribute), massaging (changing the labels of sensitive samples according to their class probability), or reweighting (giving higher weights to features from the sensitive group that are assigned a positive label) [13]. Nevertheless, from them one common approach in the literature for bias mitigation is the concept of preferential sampling, based on over- or undersampling samples from the sensitive groups with a higher probability given to borderline objects, that is, the ones with higher impact in the decision.

An example of studies that fall into this category are [31, 37].

1. Data resampling. One possible approach is to resample from the data. In [31], a negative sampler based on a KG is used to infer negative samples from the mapped positive ones, to balance the interaction data and deal with the one-class problem (e.g., only having positive feedback in e-commerce systems). This approach is using the information from an external source to complete missing information by drawing conclusions over the known sampled data, such as items close to the positive samples being negative feedback, due to a higher probability of being known by the user.
2. Data augmentation. Another approach to deal with representativeness issues in data is based on increasing the number of samples from an external source. This can be useful in situations where there are disproportions between instances, to achieve a balance between the classes. In the surveyed papers, we found an example of a statistical correction approach in [37]. Examples from a semantic resource were added to the dataset to provide it with more samples with correct annotations in order to reduce mispredictions of the model due to spurious biases in the training data.

Data enrichment. On the other hand, data enrichment methods are also based on pre-processing data to improve representativeness, but these do not increase the dataset size, but expand the features of the instances that are already part of the dataset. We found different alternatives to pre-process features, depending on which and how features are expanded.

1. Contextual enrichment. Features can be used to improve generalisation of a dataset by providing more context. Contextual enrichment is based on using additional information of the features from semantic resources to increase the feature space with properties of the features, so that content is more general and applicable in similar tasks. There exist many examples of the previous, e.g, to generalise the type of events [54–56], topics [57], and vocabulary words used in a social platform [58]. In the scope of the surveyed papers, [35] provide context to the music data to derive meaningful similarity measures that are not only popularity driven. The main problem of these methods is noise due to the introduction of a large number of uninformative features [54].
2. Sub-graph pattern mining. Another type of features which can be used are those relying on sub-graph pattern mining. Similarly to the previous, paths extracted from a representation of the features in a KG can be used to include more meaningful information into the model. For example, these semantic graph patterns have been used to identify events in text [59], or patterns of radicalisation [57]. In this survey, graph features have been useful to capture bias due to spurious input data correlations based on sensitive information or against common sense, by using association rule mining over the path patterns [46], and property specific graphs in [43] to improve similarity measures of entities that are less popular and have few data.
3. Probabilistic-based approaches. Finally, we mention probabilistic-based approaches to expand the features in neural vector representations using the knowledge extracted from a KG. These methods aim to estimate the likelihood in neural feature representations when dealing with out-of-training features. For example, KG relations can be used to estimate the weights of the features in a neural network to augment the probability of latent relations useful for automatic image captioning [32], to expand the neural feature vector recommendation with similar entities to the ones in training data [44], or to change the score of the neural network to expand the range of "detectable" words beyond the training set [38].

Data correction. Data correction approaches maintain the same number of samples and features, but change the information they represent to account for bias.

We highlight the concept of Semantic Abstraction that is applied in several studies. Using the upper-level concept of some dimensions that are not relevant to the algorithmic task (e.g., temporal, location [60, 61], or type of event), we can train more generalisable models, without dealing with the problem of noise in feature augmentation methods. In the surveyed papers, [36] is an example of abstraction of knowledge. To retract the classifier of information that is not appropriately represented in the data, these tokens are replaced by a more general token to reduce the misrepresentation and reduce adverse outcomes towards that particular concept as, for example, being incorrectly classified as a hateful post.

6.3. What challenges are encountered in the existing methods?

Although there are promising directions to improve the problem of bias in AI using semantics, we should consider the current limitations.

Evaluation variability. First, in the papers included most relevant to the current concerns of bias and discrimination in AI, semantics are mostly used to capture and mitigate bias. However, there exists a great variability in how these debiasing strategies are evaluated. There are two evaluation approaches in the studies:

1. *User-based evaluation*, which involves techniques based on the user participation in the system through experimental or observational methods, questionnaires or interviews [39, 47, 48].
2. *Baseline assessment*, which is based on using benchmark datasets, or algorithmic metrics, as a baseline or starting point, so that progress can be assessed and comparisons can be made.

Baseline automatic evaluation in the survey is generally implicit in the performance metrics of the general model. These include general ranking metrics such as recall, precision, F1, and ndcg (normalized discounted cumulative gain) of the top-K ranked items [29, 31, 43, 44], in neural networks accuracy and recall in text [37, 46] and image [30, 38] classification, or BLEU (n-gram precision) and METEOR (sentence evaluation) [32] metrics for automatic annotation tasks.

Only some studies account specifically for formal definitions of fairness metrics [29, 36, 43]. For example, in recommendation Fu et al. [29] show how different weights in the parameters of the fairness scores used to re-rank the results in terms of diversity and inequality affect the overall model performance, and [43] account for fairness more explicitly in the ranking by measuring performance in terms of serendipity and novelty. In [36], they investigate the trade-off between the accuracy of the classifier and a evaluation of the bias mitigation strategy of bias against systematic preferences in performance in different demographic groups, presumably due to challenges of fairness in society. The metric was based on a threshold-agnostic metric to provide a nuanced view on how the classifier's score distribution varied across designated groups [62]. These metrics, originally based in the $ROC-AUC$ across demographic groups and tested in a synthetically generated dataset with balanced class distributions across them, were further evaluated in a more realistic dataset in [63, 64] to provide more robust evaluations. These bias metrics have been used to audit AI systems to prevent, for example, mysogyny [65], bias against victim [66], or disability [67]) communities.

However, there is to date an ongoing discussion on how to provide not only datasets, but metrics that can be used as benchmark for bias mitigation.

- In many cases, evaluation is based on accounting for demographic information, but it should account for several form of bias in the existing models beyond those social categories that have been considered as sensitive by the US protected attributes convention [68].
- The evaluation of bias mitigation methods using semantics fall even back to this problem and require a more critical evaluation rather than only measuring the performance of the general model. This approach only captures the features that are observed in the data, the "observed" space. In principle, bias mitigation methods should aim to reduce discrepancies of this space with the features that constitute the desired basis for decision making, the "construct" space. However, this space may be not well captured leaving to the impossibility of the "fair" distribution. However, this "fair" distribution may not be well captured while being relevant to the prediction [69].

Bias within semantic resources. Secondly, we need to mention that it can not be assumed that semantic resources are free of bias, as these are prone to several limitations.

- Bias due to the data and methods that have been used to build them. For example, gender and demographic bias can seek into the automatic generated KBs due to the systematic exclusion of these groups by the neural relation extraction and named entity recognition systems that are used to automatically construct them [70, 71]. As a result, social bias can impact the downstream applications, as shown in [72] study of knowledge graph embeddings extracted from Wikidata.
- Bias due the lack of representativeness, redundancy, low coverage and noise. This has been argued in the most relevant and used resources, including DBpedia [73, 74] and Wikidata [75], Freebase [76], ConceptNet [77], WordNet [78], and FrameNet [79].

– Bias emerging at different generation stages as with general datasets. For example, we have seen an example of evaluation bias in [80], as shown in the differences between expert and crowdsource annotations in a natural language generation over knowleddge graphs task.
– Bias due to transparency and currency issues [81], as there needs to be more information about the source of the data, and whether it is up to date, in order to assess bias and ensure appropriateness before putting them into practice.

7. Conclusion

In conclusion, this survey shows the applicability of conceptual semantics to alleviate bias in AI. From over a thousand initial search results, we have followed a systematic approach and analysed 24 studies that use formal knowledge representations (i.e, taxonomies, thesaurus, ontologies, KGs, KBs, or the Linked Open Data) to identify, capture, or mitigate bias. Based on the bias and fairness literature, we provide an ample understanding and categorisation of bias, considering its type, origin, and impact.

Our findings show that semantics help in a number of AI systems (i.e., IR, RS, ML, DL, and NLP), and they do so mainly to reduce its detrimental effects or model it as part of the system to improve it. Comparing semantic methodologies to other state-of-the-art bias mitigation approaches, we see semantics fit within data augmentation (increasing the sample size), data enrichment (increasing the number of features to provide relevant information), and data correction techniques (changing the data to account for bias). Given the increasing use in the recent years of semantics, particularly of KGs, we conclude that semantics are helpful to address a significant number of biases in AI.

We bring awareness to the current limitations of previous studies, which fall mainly into the lack of understanding and need to develop more robust bias evaluation metrics that go beyond established sensitive information and space captured by the features in the dataset, as this does not necessarily capture all the relevant information necessary to build fair AI systems [69]. We also discuss some issues of bias within the semantic resources (e.g., gender and demographic bias in the data, data quality, or bias in the annotations), which need to be taken into careful consideration before putting them into practice.

However, this survey succeeds to position the work of the Semantic Web community in the past decade within the context of bias in AI and provides an analysis of the intersection of both areas of research to help future researchers identify and nurture the advantages these technologies bring. Bias in AI is a fundamental and urgent issue to ensure the applicability of automated systems in society, and semantics can be of paramount importance to improve bias accountability by extracting *sense* from the data they use.

References

[1] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernandez, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis and S. Staab, Bias in data-driven artificial intelligence systems—An introductory survey, *WIREs Data Mining and Knowledge Discovery* (2020). doi:10.1002/widm.1356.

[2] B. Hutchinson and M. Mitchell, 50 Years of Test (Un)fairness: Lessons for Machine Learning, *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19* (2019), 49–58, arXiv: 1811.10104. doi:10.1145/3287560.3287600. http://arxiv.org/abs/1811.10104.

[3] A. Romei and S. Ruggieri, A multidisciplinary survey on discrimination analysis, *The Knowledge Engineering Review* **29**(5) (2014), 582–638. doi:10.1017/S0269888913000039. https://www.cambridge.org/core/product/identifier/S0269888913000039/type/journal_article.

[4] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman and A. Galstyan, A Survey on Bias and Fairness in Machine Learning, *arXiv:1908.09635 [cs]* (2019), arXiv: 1908.09635. http://arxiv.org/abs/1908.09635.

[5] A. Olteanu, C. Castillo, F. Diaz and E. Kıcıman, Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries, *Frontiers in Big Data* **2** (2019), Publisher: Frontiers. doi:10.3389/fdata.2019.00013.

[6] R. Baeza-Yates, Bias on the web, *Communications of the ACM* **61**(6) (2018), 54–61. doi:10.1145/3209581. http://dl.acm.org/citation.cfm?doid=3229066.3209581.

[7] A. Köchling and M.C. Wehner, Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development, *Business Research* (2020). doi:10.1007/s40685-020-00134-w.

[8] A. Chouldechova, Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments, *Big Data* **5**(2) (2017), 153–163, Publisher: Mary Ann Liebert, Inc., publishers. doi:10.1089/big.2016.0047.

[9] J.R. Loftus, C. Russell, M.J. Kusner and R. Silva, Causal Reasoning for Algorithmic Fairness, *arXiv:1805.05859 [cs]* (2018), arXiv: 1805.05859. http://arxiv.org/abs/1805.05859.

[10] N. Kilbertus, M. Rojas Carulla, G. Parascandolo, M. Hardt, D. Janzing and B. Schölkopf, Avoiding Discrimination through Causal Reasoning, *Advances in Neural Information Processing Systems* **30** (2017), 656–666. https://proceedings.neurips.cc/paper/2017/hash/f5f8590cd58a54e94377e6ae2eded4d9-Abstract.html.

[11] D. Shah, H.A. Schwartz and D. Hovy, Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), 5248–5264, arXiv: 1912.11078. doi:10.18653/v1/2020.acl-main.468. http://arxiv.org/abs/1912.11078.

[12] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. Zemel, Fairness through awareness, in: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, Association for Computing Machinery, New York, NY, USA, 2012, pp. 214–226. ISBN 978-1-4503-1115-1. doi:10.1145/2090236.2090255.

[13] F. Kamiran and T. Calders, Data preprocessing techniques for classification without discrimination, *Knowledge and Information Systems* **33**(1) (2012), 1–33. doi:10.1007/s10115-011-0463-8.

[14] S.L. Blodgett, S. Barocas, H. Daumé III and H. Wallach, Language (Technology) is Power: A Critical Survey of "Bias" in NLP, *arXiv:2005.14050 [cs]* (2020), arXiv: 2005.14050. http://arxiv.org/abs/2005.14050.

[15] F. Gandon, A survey of the first 20 years of research on semantic Web and linked data, *Ingénierie des systèmes d'information* **23**(3–4) (2018), 11–38. doi:10.3166/isi.23.3-4.11-38. https://isi.revuesonline.com/article.jsp?articleId=39845.

[16] P. Brereton, B.A. Kitchenham, D. Budgen, M. Turner and M. Khalil, Lessons from applying the systematic literature review process within the software engineering domain, *Journal of Systems and Software* **80**(4) (2007), 571–583. doi:10.1016/j.jss.2006.07.009. http://www.sciencedirect.com/science/article/pii/S016412120600197X.

[17] S. Barocas and A.D. Selbst, Big Data's Disparate Impact, SSRN Scholarly Paper, ID 2477899, Social Science Research Network, Rochester, NY, 2016. doi:10.2139/ssrn.2477899. https://papers.ssrn.com/abstract=2477899.

[18] B. Lepri, N. Oliver and A. Pentland, Ethical Machines: The Human-centric Use of Artificial Intelligence, *iScience* (2021), 102249. doi:10.1016/j.isci.2021.102249. https://linkinghub.elsevier.com/retrieve/pii/S2589004221002170.

[19] A. Miles and S. Bechhofer, SKOS Simple Knowledge Organization System Reference, 2009, Publisher: World Wide Web Consortium. https://www.escholar.manchester.ac.uk/uk-ac-man-scw:66505.

[20] C. Manning and H. Schutze, *Foundations of statistical natural language processing*, MIT press, 1999.

[21] C. Fellbaum, *WordNet: An Electronic Lexical Database*, Bradford Books, 1998.

[22] T. Gruber, Ontology., *Encyclopedia of database systems* **1** (2009), 1963–1965.

[23] H. Liu and P. Singh, ConceptNet — A Practical Commonsense Reasoning Tool-Kit, *BT Technology Journal* **22**(4) (2004), 211–226. doi:http://dx.doi.org/10.1023/B:BTTJ.0000047600.45421.6d. http://portal.acm.org/citation.cfm?id=1031373.

[24] C. Bizer, T. Heath and T. Berners-Lee, Linked data: Principles and state of the art, in: *World wide web conference*, Vol. 1, 2008, p. 40.

[25] K. Petersen, S. Vakkalanka and L. Kuzniarz, Guidelines for conducting systematic mapping studies in software engineering: An update, *Information and software technology* (2015), 1–18, Publisher: Elsevier Science Section: Information and software technology. https://dialnet.unirioja.es/servlet/articulo?codigo=6000213.

[26] H.J. Lee and B.-W. Park, How to Reduce Confirmation Bias using Linked Open Data Knowledge Repository, in: *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, IEEE, Busan, Korea (South), 2020, pp. 410–416. ISBN 978-1-72816-034-4. doi:10.1109/BigComp48618.2020.00-39. https://ieeexplore.ieee.org/document/9070713/.

[27] R. Celebi, H. Uyar, E. Yasar, O. Gumus, O. Dikenelli and M. Dumontier, Evaluation of knowledge graph embedding approaches for drug-drug interaction prediction in realistic settings, *BMC Bioinformatics* **20**(1) (2019), 726. doi:10.1186/s12859-019-3284-5.

[28] F. Silva, S. Fernandes, J. Casacão, C. Libório, J. Almeida, S. Cersósimo, C.R. Mendes, R. Brandão and R. Cerqueira, Machine-Learning in Oil and Gas Exploration: A New Approach to Geological Risk Assessment, European Association of Geoscientists & Engineers, 2019, pp. 1–5, ISSN: 2214-4609 Issue: 1. doi:10.3997/2214-4609.201900988.

[29] Z. Fu, Y. Xian, R. Gao, J. Zhao, Q. Huang, Y. Ge, S. Xu, S. Geng, C. Shah, Y. Zhang and G. de Melo, Fairness-Aware Explainable Recommendation over Knowledge Graphs, in: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, Virtual Event China, 2020, pp. 69–78. ISBN 978-1-4503-8016-4. doi:10.1145/3397271.3401051.

[30] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara and S.-F. Chang, Visual Affect Around the World: A Large-scale Multilingual Visual Sentiment Ontology, in: *Proceedings of the 23rd ACM international conference on Multimedia*, MM '15, Association for Computing Machinery, New York, NY, USA, 2015, pp. 159–168. ISBN 978-1-4503-3459-4. doi:10.1145/2733373.2806246.

[31] X. Wang, Y. Xu, X. He, Y. Cao, M. Wang and T.-S. Chua, Reinforced Negative Sampling over Knowledge Graph for Recommendation, *Proceedings of The Web Conference 2020* (2020), 99–109, arXiv: 2003.05753. doi:10.1145/3366423.3380098. http://arxiv.org/abs/2003.05753.

[32] F. Huang, Z. Li, S. Chen, C. Zhang and H. Ma, Image Captioning with Internal and External Knowledge, in: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 535–544. ISBN 978-1-4503-6859-9. doi:10.1145/3340531.3411948.

[33] H. Chen, A. Trouve, K.J. Murakami and A. Fukuda, Semantic image retrieval for complex queries using a knowledge parser, *Multimedia Tools and Applications* **77**(9) (2018), 10733–10751. doi:10.1007/s11042-017-4932-2.

[34] H.-J. Kim and M. Song, An Ontology-Based Approach to Sentiment Classification of Mixed Opinions in Online Restaurant Reviews, in: *Social Informatics*, A. Jatowt, E.-P. Lim, Y. Ding, A. Miura, T. Tezuka, G. Dias, K. Tanaka, A. Flanagin and B.T. Dai, eds, Lecture Notes in Computer Science, Springer

International Publishing, Cham, 2013, pp. 95–108. ISBN 978-3-319-03260-3. doi:10.1007/978-3-319-03260-3$_9$.

[35] G.K. Koduri, Culture-Aware Approaches to Modeling and Description of Intonation Using Multimodal Data, in: *Knowledge Engineering and Knowledge Management*, P. Lambrix, E. Hyvönen, E. Blomqvist, V. Presutti, G. Qi, U. Sattler, Y. Ding and C. Ghidini, eds, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2015, pp. 209–217. ISBN 978-3-319-17966-7. doi:10.1007/978-3-319-17966-7$_3$0.

[36] P. Badjatiya, M. Gupta and V. Varma, Stereotypical Bias Removal for Hate Speech Detection Task using Knowledge-based Generalizations, in: *The World Wide Web Conference*, WWW '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 49–59. ISBN 978-1-4503-6674-8. doi:10.1145/3308558.3313504.

[37] M. Mensio, E. Bastianelli, I. Tiddi and G. Rizzo, Mitigating bias in deep nets with knowledge bases: The case of natural language understanding for robots, in: *CEUR Workshop Proceedings*, Vol. 2600, CEUR-WS, 2020, p. 20, ISSN: 1613-0073. https://researchportal.hw.ac.uk/en/publications/mitigating-bias-in-deep-nets-with-knowledge-bases-the-case-of-nat.

[38] R.T. Icarte, J.A. Baier, C. Ruz and A. Soto, How a General-Purpose Commonsense Ontology can Improve Performance of Learning-Based Image Retrieval (2017). https://arxiv-org.libezproxy.open.ac.uk/abs/1705.08844v1.

[39] R. Ludolph, A. Allam and P.J. Schulz, Manipulating Google's Knowledge Graph Box to Counter Biased Information Processing During an Online Search on Vaccination: Application of a Technological Debiasing Strategy, *Journal of Medical Internet Research* **18**(6) (2016), e137. doi:10.2196/jmir.5430. http://www.jmir.org/2016/6/e137/.

[40] E. Sediyono, Suhartono and C. Nivak, Measuring the Performance of Ontological Based Information Retrieval from a Social Media, in: *2014 European Modelling Symposium*, 2014, pp. 354–359. doi:10.1109/EMS.2014.15.

[41] C.-E. Wu and R.T.-H. Tsai, Using relation selection to improve value propagation in a ConceptNet-based sentiment dictionary, *Knowledge-Based Systems* **69** (2014), 100–107. doi:10.1016/j.knosys.2014.04.043. http://www.sciencedirect.com/science/article/pii/S0950705114001737.

[42] A.L. Garrido, M.G. Buey, G. Muñoz and J.-L. Casado-Rubio, Information Extraction on Weather Forecasts with Semantic Technologies, in: *Natural Language Processing and Information Systems*, E. Métais, F. Meziane, M. Saraee, V. Sugumaran and S. Vadera, eds, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2016, pp. 140–151. ISBN 978-3-319-41754-7. doi:10.1007/978-3-319-41754-7$_1$2.

[43] E. Palumbo, D. Monti, G. Rizzo, R. Troncy and E. Baralis, entity2rec: Property-specific knowledge graph embeddings for item recommendation, *Expert Systems with Applications* **151** (2020), 113235. doi:10.1016/j.eswa.2020.113235. http://www.sciencedirect.com/science/article/pii/S0957417420300610.

[44] Q. Chen, J. Lin, Y. Zhang, M. Ding, Y. Cen, H. Yang and J. Tang, Towards Knowledge-Based Recommender Dialog System, 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)*, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1803–1813. doi:10.18653/v1/D19-1189. https://www.aclweb.org/anthology/D19-1189.

[45] R. Awadallah, M. Ramanath and G. Weikum, OpinioNetIt: A Structured and Faceted Knowledge-base of Opinions, in: *2012 IEEE 12th International Conference on Data Mining Workshops*, 2012, pp. 878–881, ISSN: 2375-9259. doi:10.1109/ICDMW.2012.49.

[46] A. Nikolov and M. d'Aquin, Uncovering Semantic Bias in Neural Network Models Using a Knowledge Graph, in: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1175–1184. ISBN 978-1-4503-6859-9. doi:10.1145/3340531.3412009.

[47] L. Gregori, R. Varvara and A.A. Ravelli, Action Type induction from multilingual lexical features, *Procesamiento del Lenguaje Natural* **63**(0) (2019), 85–92, Number: 0. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6098.

[48] B. Sarrafzadeh, A. Vtyurina, E. Lank and O. Vechtomova, Knowledge Graphs versus Hierarchies: An Analysis of User Behaviours and Perspectives in Information Seeking, in: *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, Association for Computing Machinery, New York, NY, USA, 2016, pp. 91–100. ISBN 978-1-4503-3751-9. doi:10.1145/2854946.2854958.

[49] S. Chelaru, I.S. Altingovde, S. Siersdorfer and W. Nejdl, Analyzing, Detecting, and Exploiting Sentiment in Web Queries, *ACM Transactions on the Web* **8**(1) (2013), 6:1–6:28. doi:10.1145/2535525.

[50] G. Tecuci, D. Schum, D. Marcu and M. Boicu, Recognizing and countering biases in intelligence analysis with TIACRITIS, *CEUR Workshop Proceedings* **1097** (2013), 25–32.

[51] G. Lortal, P. Capet and A. Bertone, Ontology building for cognitive bias assessment in intelligence, in: *2014 IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, 2014, pp. 237–243, ISSN: 2379-1675. doi:10.1109/CogSIMA.2014.6816616.

[52] A. Chatterjee, S. Joshi, P. Bhattacharyya, D. Kanojia and A. Meena, A Study of the Sense Annotation Process: Man v/s Machine, 2012.

[53] R.J. Heuer, *Psychology of Intelligence Analysis*, Center for the Study of Intelligence, 1999, Google-Books-ID: rRXFhKAiG8gC. ISBN 978-1-929667-00-0.

[54] S. Romero and K. Becker, Improving the classification of events in tweets using semantic enrichment, in: *Proceedings of the International Conference on Web Intelligence*, WI '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 581–588. ISBN 978-1-4503-4951-2. doi:10.1145/3106426.3106435.

[55] P. Khare, G. Burel and H. Alani, Classifying Crises-Information Relevancy with Semantics, in: *The Semantic Web*, Vol. 10843, A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai and M. Alam, eds, Springer International Publishing, Cham, 2018, pp. 367–383, Series Title: Lecture Notes in Computer Science. ISBN 978-3-319-93416-7 978-3-319-93417-4. doi:10.1007/978-3-319-93417-4$_2$4.

[56] S. Romero and K. Becker, A framework for event classification in tweets based on hybrid semantic enrichment, *Expert Systems with Applications* **118** (2019), 522–538. doi:10.1016/j.eswa.2018.10.028. http://www.sciencedirect.com/science/article/pii/S095741741830678X.

[57] B.L. Ibtihel, H. Lobna and B.J. Maher, A Semantic Approach for Tweet Categorization, *Procedia Computer Science* **126** (2018), 335–344. doi:10.1016/j.procs.2018.07.267. https://www.sciencedirect.com/science/article/pii/S1877050918312432.

[58] M. Grzeça, K. Becker and R. Galante, Improving the Classification of Drunk Texting in Tweets Using Semantic Enrichment, in: *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 2018, pp. 190–197. doi:10.1109/WI.2018.00-90.

[59] T. Dickinson, M. Fernandez, L.A. Thomas, P. Mulholland, P. Briggs and H. Alani, Identifying Important Life Events from Twitter Using Semantic and Syntactic Patterns, in: *WWW/Internet Conference proceedings 2016*, IADIS Press, Mannheim, Germany, 2016, pp. 143–150. ISBN 978-989-8533-57-9. http://oro.open.ac.uk/48679/.

[60] A. Schulz and F. Janssen, What is good for one city may not be good for another one: evaluating generalization for tweet classification based on semantic abstraction, in: *Proceedings of the Fifth International Conference on Semantics for Smarter Cities - Volume 1280*, S4SC'14, CEUR-WS.org, Aachen, DEU, 2014, pp. 53–67.

[61] A. Schulz, C. Guckelsberger and F. Janssen, Semantic Abstraction for generalization of tweet classification: An evaluation of incident-related tweets, *Semantic Web* **8**(3) (2017), 353–372, Publisher: IOS Press. doi:10.3233/SW-150188. https://content-iospress-com.libezproxy.open.ac.uk/articles/semantic-web/sw188.

[62] L. Dixon, J. Li, J. Sorensen, N. Thain and L. Vasserman, Measuring and Mitigating Unintended Bias in Text Classification, in: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 67–73. ISBN 978-1-4503-6012-8. doi:10.1145/3278721.3278729.

[63] D. Borkan, L. Dixon, J. Li, J. Sorensen, N. Thain and L. Vasserman, Limitations of Pinned AUC for Measuring Unintended Bias, *arXiv:1903.02088 [cs, stat]* (2019), arXiv: 1903.02088. http://arxiv.org/abs/1903.02088.

[64] D. Borkan, L. Dixon, J. Sorensen, N. Thain and L. Vasserman, Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification (2019). https://arxiv-org.libezproxy.open.ac.uk/abs/1903.04561v2.

[65] D. Nozza, C. Volpetti and E. Fersini, Unintended Bias in Misogyny Detection, in: *IEEE/WIC/ACM International Conference on Web Intelligence*, WI '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 149–155. ISBN 978-1-4503-6934-3. doi:10.1145/3350546.3352512.

[66] B. Mathew, P. Saha, S.M. Yimam, C. Biemann, P. Goyal and A. Mukherjee, HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection (2020). https://arxiv-org.libezproxy.open.ac.uk/abs/2012.10289v1.

[67] B. Hutchinson, V. Prabhakaran, E. Denton, K. Webster, Y. Zhong and S. Denuyl, Social Biases in NLP Models as Barriers for Persons with Disabilities, *arXiv:2005.00813 [cs]* (2020), arXiv: 2005.00813. http://arxiv.org/abs/2005.00813.

[68] E.M. Bender, T. Gebru, A. McMillan-Major and S. Shmitchell, On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? , in: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ACM, Virtual Event Canada, 2021, pp. 610–623. ISBN 978-1-4503-8309-7. doi:10.1145/3442188.3445922.

[69] S.A. Friedler, C. Scheidegger and S. Venkatasubramanian, On the (im)possibility of fairness, *arXiv:1609.07236 [cs, stat]* (2016), arXiv: 1609.07236. http://arxiv.org/abs/1609.07236.

[70] A. Gaut, T. Sun, S. Tang, Y. Huang, J. Qian, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang and W.Y. Wang, Towards Understanding Gender Bias in Relation Extraction, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, 2020, pp. 2943–2953. doi:10.18653/v1/2020.acl-main.265. https://www.aclweb.org/anthology/2020.acl-main.265.

[71] S. Mishra, S. He and L. Belli, Assessing Demographic Bias in Named Entity Recognition, *arXiv:2008.03415 [cs]* (2020), arXiv: 2008.03415. http://arxiv.org/abs/2008.03415.

[72] J. Fisher, D. Palfrey, C. Christodoulopoulos and A. Mittal, Measuring Social Bias in Knowledge Graph Embeddings, *arXiv:1912.02761 [cs]* (2020), arXiv: 1912.02761. http://arxiv.org/abs/1912.02761.

[73] A. Soulet, A. Giacometti, B. Markhoff and F.M. Suchanek, Representativeness of Knowledge Bases with the Generalized Benford's Law, in: *The Semantic Web – ISWC 2018*, D. Vrandečić, K. Bontcheva, M.C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L.-A. Kaffee and E. Simperl, eds, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2018, pp. 374–390. ISBN 978-3-030-00671-6. doi:10.1007/978-3-030-00671-$6_2$2.

[74] K. Janowicz, B. Yan, B. Regalia, R. Zhu and G. Mai, Debiasing knowledge graphs: Why Female Presidents are not like Female Popes, 2018.

[75] B. Kruit, P. Boncz and J. Urbani, Extracting Novel Facts from Tables for Knowledge Graph Completion, in: *The Semantic Web – ISWC 2019*, C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I. Cruz, A. Hogan, J. Song, M. Lefrançois and F. Gandon, eds, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2019, pp. 364–381. ISBN 978-3-030-30793-6. doi:10.1007/978-3-030-30793-$6_2$1.

[76] X. Wang, X. Han, Z. Liu, M. Sun and P. Li, Adversarial Training for Weakly Supervised Event Detection, 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)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 998–1008. doi:10.18653/v1/N19-1105. https://www.aclweb.org/anthology/N19-1105.

[77] J. Feldman, J. Davison and A.M. Rush, Commonsense Knowledge Mining from Pretrained Models, *arXiv:1909.00505 [cs]* (2019), arXiv: 1909.00505. http://arxiv.org/abs/1909.00505.

[78] J. Niu, Z. Sun and W. Zhang, Enhancing Knowledge Graph Completion with Positive Unlabeled Learning, in: *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 296–301, ISSN: 1051-4651. doi:10.1109/ICPR.2018.8545743.

[79] S. Tonelli, C. Giuliano and K. Tymoshenko, Wikipedia-based WSD for multilingual frame annotation, *Artificial Intelligence* **194** (2013), 203–221. doi:10.1016/j.artint.2012.06.002.

https://www.sciencedirect.com/science/article/pii/S0004370212000720.

[80] P. Vougiouklis, E. Maddalena, J. Hare and E. Simperl, How Biased Is Your NLG Evaluation? (short paper), in: *SAD/CrowdBias@HCOMP*, 2018.

[81] C.T. Wolf, From Knowledge Graphs to Knowledge Practices: On the Need for Transparency and Explainability in Enterprise Knowledge Graph Applications (2020), 3.

[82] A. Kutuzov, L. Øvrelid, T. Szymanski and E. Velldal, Diachronic word embeddings and semantic shifts: a survey (2018), 14.

[83] A. Lauscher, G. Glavaš, S.P. Ponzetto and I. Vulić, A General Framework for Implicit and Explicit Debiasing of Distributional Word Vector Spaces, *arXiv:1909.06092 [cs]* (2020), arXiv: 1909.06092. http://arxiv.org/abs/1909.06092.

[84] M. Atzmueller, Declarative Aspects in Explicative Data Mining for Computational Sensemaking, in: *Declarative Programming and Knowledge Management*, Vol. 10997, D. Seipel, M. Hanus and S. Abreu, eds, Springer International Publishing, Cham, 2018, pp. 97–114, Series Title: Lecture Notes in Computer Science. ISBN 978-3-030-00800-0 978-3-030-00801-7. doi:10.1007/978-3-030-00801-7$_7$.

[85] S. Halford, J.A. Hendler, E. Ntoutsi and S. Staab, 10 Years of Web Science: Closing The Loop (Dagstuhl Perspectives Workshop 18262) (2019), 26 pages, Artwork Size: 26 pages Medium: application/pdf Publisher: Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik GmbH, Wadern/Saarbruecken, Germany Version Number: 1.0. doi:10.4230/DAGREP.8.6.173. http://drops.dagstuhl.de/opus/volltexte/2019/10059/.

[86] H. Hedden, Fast Track to Knowledge Graphs and Semantic AI - Module 1, 94.

[87] M. Cunneen, M. Mullins, F. Murphy and S. Gaines, Artificial Driving Intelligence and Moral Agency: Examining the Decision Ontology of Unavoidable Road Traffic Accidents through the Prism of the Trolley Dilemma, *Applied Artificial Intelligence* **33**(3) (2019), 267–293, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/08839514.2018.1560124. doi:10.1080/08839514.2018.1560124.

[88] M. Senda, D. Iwasa, T. Hayashi and Y. Ohsawa, Data Classification by Reducing Bias of Domain-Oriented Knowledge Based on Data Jackets, in: *2019 6th International Conference on Signal Processing and Integrated Networks (SPIN)*, IEEE, Noida, India, 2019, pp. 404–407. ISBN 978-1-72811-380-7. doi:10.1109/SPIN.2019.8711715. https://ieeexplore.ieee.org/document/8711715/.

[89] J. Euzenat, M.-E. Roşoiu and C. Trojahn, Ontology matching benchmarks: Generation, stability, and discriminability, *Journal of Web Semantics* **21** (2013), 30–48. doi:10.1016/j.websem.2013.05.002. http://www.sciencedirect.com/science/article/pii/S1570826813000188.

[90] F. Croce, G. Cima, M. Lenzerini and T. Catarci, Ontology-based explanation of classifiers, in: *EDBT/ICDT Workshops*, 2020.

[91] H. Saif, T. Dickinson, L. Kastler, M. Fernandez and H. Alani, A Semantic Graph-Based Approach for Radicalisation Detection on Social Media, in: *Lecture Notes in Computer Science*, Vol. 10249, Springer, Portorož, Slovenia, 2017, pp. 571–587, ISSN: 0302-9743. http://oro.open.ac.uk/49640/.