Semantic Web 14 (2023) 745–770
DOI 10.3233/SW-223041
IOS Press

Semantic Web technologies and bias in artificial intelligence: A systematic literature review

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

Editor: Dagmar Gromann, University of Vienna, Austria
Solicited reviews: Konstantinos Kotis, University of the Aegean, Greece; Dagmar Gromann, University of Vienna, Austria; one anonymous reviewer

Abstract. Bias in Artificial Intelligence (AI) is a critical and timely issue due to its sociological, economic and legal impact, as decisions made by biased algorithms could lead to unfair treatment of specific individuals or groups. Multiple surveys have emerged to provide a multidisciplinary view of bias or to review bias in specific areas such as social sciences, business research, criminal justice, or data mining. Given the ability of Semantic Web (SW) technologies to support multiple AI systems, we review the extent to which semantics can be a "tool" to address bias in different algorithmic scenarios. We provide an in-depth categorisation and analysis of bias assessment, representation, and mitigation approaches that use SW technologies. We discuss their potential in dealing with issues such as representing disparities of specific demographics or reducing data drifts, sparsity, and missing values. We find research works on AI bias that apply semantics mainly in information retrieval, recommendation and natural language processing applications and argue through multiple use cases that semantics can help deal with technical, sociological, and psychological challenges.

Keywords: Bias in Artificial Intelligence, Semantic Web technologies, bias assessment, bias representation, bias mitigation, algorithmic fairness

1. Introduction

There is growing awareness of bias and discrimination in AI applications. Users from inactive groups are more at risk of being mistreated on e-commerce platforms, such as Amazon or eBay, which is problematic as these often correspond to limited income groups [29]. One of the main challenges in image searching is its limitation to only the sample set of training data [19,41], which can lead to irrelevant or inaccurate results but, at worst, incorrect associations that reflect and perpetuate the harm done to historically disadvantaged groups [77]. These are just a few enlightening examples of the use cases covered in this survey article. Understandably, the direction of the AI community is shifting towards the pursuit of not only accurate but also ethical AI [50].

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

One of the main advantages of AI over human intelligence is its ability to process vast amounts of data. Indeed, data plays a fundamental role in which algorithmic decisions can reproduce or even amplify human biases, as these systems are only as good as the data they work with [9]. One of the main challenges of AI is dealing with data limitations, such as incomplete, unrepresentative and erroneous data [7]. In addition, how humans do or do not have access to these systems and how they interact with them are also key bias factors to consider in AI design.

The vast amount of information available on the Semantic Web (SW) has enormous potential to address the bias problems mentioned above by leveraging the structured formalisation of machine-understandable knowledge to build more realistic and fairer models. There are examples in different domains, such as machine learning and data mining, natural language processing or social networking and media representation, where the SW, linked data and the web of data have made a significant contribution [30]. For example, we can leverage semantics to control and restrict personal and sensitive data access, support different AI processing tasks such as reasoning, mining, clustering and learning, or extract arguments from natural language text. These systems are not averse to systemic bias, e.g., they may lack information from specific domains or entities that are less popular than others, or information from specific demographics may have more detail depending on the contributor's interest [42]. While we consider the potential bias that SW technologies may have at the data and schema levels, we mainly focus on the contribution that SW technologies can make as a "tool" to address bias in different algorithmic scenarios to promote algorithmic fairness. This analysis is relevant for the SW and AI communities, as bias is gaining attention in different areas, such as computer science, social sciences, philosophy and law [50].

In this article, we provide a review of the contribution of SW technologies to addressing bias in AI. We aim to explain why bias arises and at what level of the AI system, *to better understand harmful behaviours*, and how bias manifests *to understand better whom it affects and how*. This in-depth conceptualisation is crucial due to the lack of consistency between the motivation, and the technological solutions proposed to address bias in AI [12], as we need to understand what system behaviours are considered harmful, in what way, to whom and why.

We follow a systematic approach [13] to review the literature and analyse the existing bias solutions that use semantics. Specifically, we focus on the following contributions:

i) We provide a survey of 34 papers that use semantic-based techniques to address bias in AI.
ii) We categorise relevant papers according to the type of semantics used, and the type of bias they target.
iii) We highlight the most common AI application areas in the framework of semantic research for bias.
iv) We identify further challenges in AI bias research for the SW and AI communities.

The rest of the paper is organised as follows. In Section 2, we describe the methodology of the systematic literature review. In Section 3, we define the concepts of semantics and bias used in this article. In Section 4, we report on an analysis of previous works that use semantics to address bias in AI and discuss the main findings, future opportunities and challenges in Section 5. Finally, we provide a conclusion in Section 6.

2. Survey methodology

To provide a thorough literature review, we followed the guidelines of the systematic mapping study research method [13]. Specifically, we address the following research question (RQ):

To what extent can SW technologies be used to address bias in AI?

Two main components constitute this RQ: *semantics* and *bias*. The first aims to investigate the SW community's contribution in methods, evaluation frameworks, or metrics to address bias in AI. The second focuses on bias, aiming to assess the types and sources of bias that semantic knowledge can address and the main challenges in AI that semantics can help overcome.

The collection of relevant works is based on keyword-based querying in two popular scholarly databases: Elsevier Scopus and ISI Web of Knowledge (WoS) (Table 1). We complete our search with Microsoft Academic Search, Semantic Scholar and Google Scholar to do snowballing [68]. We collect papers according to specific inclusion criteria (IC):

IC1: Papers written in English.

Table 1
Keywords used to search for relevant works in scholarly databases. TITLE-ABS-KEY refer to the title, abstract and keywords of the paper, respectively. We use the wildcard * to ensure multiple spelling variations are included in the search results

Search keywords
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*')

IC2: Papers published in relevant journals between 2010 and 2021.
IC3: Only papers subjected to peer review, including published journal papers, as part of conference proceedings or workshops and book chapters.

A list of venues representative of the papers found includes the International Semantic Web Conference (ISWC), the European Semantic Web Conference (ESWC), the World Wide Web Conference (WWW), the International Conference on Information and Knowledge Management (CIKM), the Conference on Artificial Intelligence (AAAI), the International Joint Conference on Artificial Intelligence (IJCAI), and the Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Two reviewers filtered the papers in four subsequent steps (Fig. 1).

In the *Source-based* filter, we select Computer Science, Mathematics, Engineering, Business, Decision Science and Social Sciences as relevant sources when using Scopus. In WoS, we consider all search results and use them to complete our search by adding all non-duplicate articles to our list of related papers.

The *Metadata-based* filter is a paper screening based on title, abstract, publication venue and publication year to discard papers not relevant to our RQ. We consider project proposals and literature reviews in the discussion, but not as *Use Cases* in our analysis. In the case of papers published in more than one venue, we include their latest version.

The *Content-based* filter consists of a paper screening based on the introduction, conclusion, or full text, especially in unclear studies. This research paper aims to investigate the SW technologies used in solutions for bias coming from the use or development of AI. Therefore, we exclude papers that lack an AI system or the use of SW technologies. Some works lack evidence of improving bias, e.g., there is no experiment or vision on how to address bias in recommendation systems [49]. Others do not use semantics in their solution, e.g., they use knowledge graph embeddings but addresses bias using disjoint test classes [15].

The *Snowballing* process concludes our search by including additional studies from paper citations when reading the filtered papers in more detail. Out of the identified 58 relevant works, 34 are examples of *Use Cases* that use semantics to address bias. The remaining 24 are surveys, position papers, and research works addressing bias within semantic resources relevant to the discussion.

3. Semantics to address bias in AI

This section presents the definitions for semantics and the conceptualisations of bias in AI used for this paper's analysis.

3.1. Semantics

There are various SW technologies (e.g. taxonomies, thesauri, ontologies, or knowledge graphs). This section defines the specific semantic resources in the surveyed papers to support a better understanding.

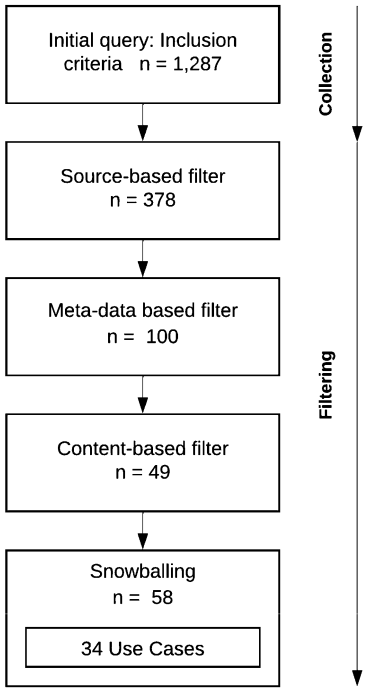

Fig. 1. Filtering relevant works of semantics to address bias in AI.

Lexical resources are representations of general language in a relational structure [70]. They have standard structured relationships and other properties for each concept, such as related and alternative terms. An example of this type of resource is WordNet [58], a commonly known lexical database of English.

An ontology defines a set of classes, attributes and relationships that model a knowledge domain with varying levels of expressivity [35]. For example, these formal and explicit specifications can be of a shared conceptualisation of meteorological variables (temperature, precipitation, visibility) to capture the domain of weather forecasting [31], the feelings and emotions conveyed by visual features to capture the psychology of human affect [43], or the reasoning steps of tasks involving problem-solving in specific domains (writing a risk assessment in industrial, insurance, health or environmental domains [53]). In particular, these forms of *knowledge representation* can capture terms or statements about the real world at different levels of domain specialisation [73]. This scope gives rise to foundational or top-level, general or core-reference ontologies. The most commonly used in the surveyed papers are domain ontologies (e.g., for the travel and tourism domain [76]) and application ontologies (e.g., for the extraction of information from a weather forecast written in natural language [31]).

A knowledge graph (KG) is a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent potentially different relations between these entities [37]. This form of data representation as a graph represents concepts, classes, properties, relationships and entity descriptions. ConceptNet [52] is an example of a KG of 1.6 million assertions of commonsense knowledge. A statement such as "cooking food can be fun" can be represented in the graph as `<cook food> <capableOf> <be fun>`. Some common applications of KGs include recommendation engines, question answering, or enterprise knowledge management [89]. Examples of popular open-source KGs are DBpedia [4] and Wikidata [84].

Finally, some works propose solutions based on the Linked Open Data (LOD), in which all SW technologies can be represented. LOD refers to a set of best practices for publishing and connecting structured data on the Web [11]. LOD relies on documents containing data in RDF (Resource Description Framework) format to make links between arbitrary *things in the world*, i.e., typed hyperlinks to the related entities in other data sources. Therefore, instead of navigating between web pages, Linked Data browsers allow users to navigate between data sources connected by specific entities. For example, the LOD platform of the Open University allows users to navigate the

University's content (courses or scholarly publications) and establish connections with other educational institutions [23]. Typically, KGs are published following the linked data principles.

This survey aims to capture the role SW technologies such as those defined above play in addressing bias in AI. Specifically, we find research works that use semantics for three high-level tasks:

Assessing bias Semantics can uncover bias. As an example, the representation of user-item interactions as a graph is used in [29] to discover disparities in the quality of recommendations in user groups with less historical data. There is great value in uncovering inequalities through observable user properties in algorithmic scenarios where information about groups vulnerable to discriminatory treatment is unknown.

Representing bias Semantics can capture bias to make it explicit and raise awareness of its implications. The use of semantic representations can help to include information about underrepresented groups in the data, e.g., due to lack of linguistic coverage in a dataset for visual sentiment prediction [43]. Documenting consistent errors in black-box models [61] or humans when using such systems [80] can help prevent them and take action.

Mitigating bias Semantics can reduce the negative impact of bias in AI systems. Therefore, we investigate the combination of SW technologies with bias mitigation methods. Bias mitigation has generally been divided into three groups [56,62]: those focusing on changing the training data [2,6,19,21,24,39,41,46,47,57,85], the learning algorithm during the model generation [3,31,51,54,76,88], or the model outcomes according to the results in a holdout dataset which was not involved during the training phase [29]. Such methods may mitigate undesirable associations of specific demographic groups with hateful connotations. For example, to prevent sentences like "he is gay" or "one of John brothers was homosexual while the other is a black transgender" from having high toxicity scores [6].

3.2. Bias in AI

We aim to help the understanding of bias in AI through examples that explain why it can occur and where it comes from, as this knowledge is necessary to understand the analysis of the harmful effects and impact of bias in Section 4. Bias can be defined as heterogeneities in data due to being generated by subgroups of people with their own characteristics and behaviours [56]. A model learned from biased data may lead to unfair and inaccurate predictions. Furthermore, bias can lead to unfairness due to systematic errors made by algorithms that lead to adverse or undesired outcomes, for example, of a particular group that has been historically disadvantaged [9,62]. This example of discriminatory behaviour is particularly concerning since AI systems have proven to reproduce or even amplify inequalities in society.

Table 2 presents the surveyed papers attending to the possible nature of the errors [7]. From a psychological perspective, systematic errors can occur due to the way humans process and interpret information and constitute a *cognitive* bias, which has shown to affect all decision-making steps [22,53,80]. In web search, this can lead to impaired judgement due to the human's heuristic way of processing information [54,74] and, in more severe cases, to group polarisation [18,67]. In machine learning (ML), the absence of the context about the domain of the text has shown to leave annotators in an indecisive state so that their annotations incorrectly shift towards the most frequent sense of a word [17]. The use of subjective text and opinions or any data arising from human interpretation can also have challenging impacts if used to develop AI applications [3,31,46,51].

From a statistical point of view, systematic deviations of the, possibly unknown, real distributions of the variables represented in the data can lead to inaccurate estimations and constitute a *statistical* bias. For example, representation disparities in the data of the users [29], items [2,20,65], or their recorded interactions [85] can compromise the

Table 2
Categories of bias attending to the nature of the errors [7]

Type of bias	#Papers	Reference
Cognitive bias	12	[3,17,18,22,31,46,51,53,54,67,74,80]
Statistical bias	16	[2,14,19–21,24,27,29,38,39,41,57,65,76,85,88]
Cultural bias	6	[5,6,34,43,47,61]

Table 3
Categories of bias depending on the location in the AI workflow where bias originates [64]

Bias location	Due to	Reference
Bias at source	External bias	[5,6,18,61,67]
	Functional bias	[2,27,29,54,65,74,85]
Bias at collection	Sampling	[34,43,47]
	Querying	[19,76]
Data pre-processing	Annotation	[3,14,17,20,21,24,31,38,39,41,46,51,57]
	Aggregation	[88]
Data analysis	Inference and prediction	[22,53,80]

quality and fairness of RS. Searching for information based only on the distributions of a specific dataset can lead to irrelevant results or results biased to other meanings of the words used in the query [19,76]. Similarly, the use of small, domain-specific datasets for training black-box models can lead to undesired behaviours, such as missing the image objects needed to provide meaningful captions [39] and answers to a question about the image [21], or retrieve relevant results to a search query [41], or missing the correct words that enable robots to understand the commands given in a sentence [57]. Consequently, predictions based on these datasets may lead to making decisions based on correlations that are unacceptable in specific cases, as these data only provide estimates from limited settings (e.g. randomised controlled trials [38] or specific datasets commonly used as benchmarks [14]). Such data limitation is especially concerning in clinical research, as there is a risk of exclusion of women and minorities [27].

From a sociological perspective, data may contain existing biases and beliefs that reflect historical and social inequalities, which the AI system may learn. It constitutes what is known as *cultural* or historical bias. The lack of diversity and overrepresentation of commercial music may lead to music recommendation platforms that are biased towards the specific cultures of popular music [47]. The use of a predominant language, such as English, can lead to generalist systems that are not inclusive of other cultures, for example, when retrieving images based on the feelings they evoke [43] or videos showing what a particular action entails [34]. In some cases, generalised beliefs about particular groups of people are reflected in the data and can lead to learning incorrect associations of these groups with undesirable attributes [6,61].

This categorisation is crucial to reveal how semantics can help with data (i.e. statistical, cultural) and user-dependent (i.e. cognitive) biases. Another aspect to consider is where in the AI workflow these biases originate, so we rely on the framework in [64] for the domain of social media analysis as it closely reflects general practices in AI. We give this general overview as the description in the semantics-specific framework of Section 4 discusses the works using more concrete concepts of bias (bias at a lower level).

We use the examples in the surveyed papers shown in Table 3. The first critical point of bias is at the data origin or source since any bias existing at the input of an AI system will appear at least in the same way at the output (i.e. "garbage in – garbage out" principle). This is the bias origin most predominant in the surveyed papers, particularly due to *external* or *functional* factors. The first concerns factors outside the AI system that can influence the reliability and representativeness of the data. For example, the prejudice against specific demographic groups [6,61], or context of a specific political affiliation [5], or community views about particular topics [18,67] may be reflected in the dataset and limit the generalisability of the conclusions that can be drawn from it. The second involves similar limitations due to the design of the AI system. For example, using only purchase data [29], positive feedback [85], or popular items [2,65] for recommendation affects the data usability. Specific designs can shape and condition users' behaviours, e.g., the ranking of search results influences the quality of the information gathered [54,74]. The heterogeneity of platforms may impede the identification of phenomena analysed on a large scale and also limits the treatment of biases in different study settings [27].

Data collection is the second step in which bias can appear, and examples of this type found in the surveyed papers include *sampling* and *querying* bias. Sampling bias occurs when the data sample is not representative of the whole population, e.g., is only collected from the most popular sources [47] or language [34,43], so the data collected is not representative of minority groups. Querying bias may emerge due to the lack of expressiveness in the possible

query formulations to be able to search for the necessary information, e.g., in an image [19] or information [76] retrieval systems.

Data pre-processing is susceptible to bias, in particular in this study, of *annotation* and *aggregation*. Noisy labels due to poor or missing guidelines compromise manual annotations (of meteorological analytical data [31], a product review [46], the meaning [17] or link between words in a text [3], or the description of abnormalities in medical images [51]), and frequently lead to the use of small corpora which cannot generalise to novel examples [14,20, 21,39,41,57]. In domains or problems where the ground truth may not be well defined (e.g., making a medical diagnosis), the use of annotated corpora has limited capacity to ensure that human experts reach a specific level of understanding so that these systems can be applied effectively, efficiently and satisfactorily [38]. An example of bias when transforming the data to infer new facts is found in [88], where bias arises due to an imbalance of the two classes used to infer new sentiment values.

Data analysis is the last step covered by this study that can cause bias, specifically at *inference and prediction* time. For example, issues of this sort may arise when using data as a source of hypotheses rather than a tool to test them [80] or making consistent and predictable mistakes during intelligence activity tasks that draw conclusions from data [22,53].

4. Description of approaches

Our analysis aims to understand better how bias impacts different AI systems and provides specific methodological examples that can apply to similar problems in future research. This section provides the review of bias assessment, representation and mitigation methodologies that use semantics, which we present following the order in Table 4.

4.1. Semantics to assess bias

Assessing biases is a fundamental task in analysing and interpreting model behaviours. It can reveal intrinsic biases that are difficult to detect due to the opaque nature of many AI systems [57]. The following examples of works are presented as semantics use cases to help with this problem.

4.1.1. Bias affecting specific groups of people

KGs can help assess recommendation disparities in user groups that are less active [29] (e.g. economically disadvantaged users). This problem constitutes a *population bias* because it affects groups that are underrepresented in the data with respect to the most active users. Therefore, their historic user-interaction data is less visible in the recommendation system. The harmful effects of this bias impact the recommendation quality and diversity of the results, posing a fairness problem. A fairness-aware algorithm that leverages entities, relationships and paths in KGs is proposed to explicitly model the recommendations in reasoning paths and apply constraints that impose fairness across users. In this case, the Amazon item e-commerce KG of entities and relations is used to quantify the richness and evenness of recommendations, revealing disparities of groups with historically less user-item interaction data. This semantic knowledge is crucial since it is a setting where users do not disclose the personal information required to deal with possible discriminatory treatments (i.e. sensitive features such as gender, age, or religion). Richness and evenness disparities are measured with each user's number of graph patterns and the relative importance of each pattern across users, respectively. The paper shows how these measures can also be used as fairness constraints to improve the quality and diversity of recommendations for these vulnerable user groups.

The under-representation of certain demographic groups was assessed in drug exposure studies. Due to the heterogeneity and lack of metadata in the gene expression databases resulting from these medical studies, it is challenging to examine large-scale differences in sex representation of the data. This assessment is essential, as women are 50% more likely to suffer from adverse drug effects. In [27], they were able to assess sex bias in public repositories of biological data by mapping existing and inferred metadata (using ML models) to existing medical ontologies. Specifically using Cellosaurus and DrugBank databases to identify cell lines and drugs, respectively. They could label drug studies from all publicly available samples using named entity recognition to identify drug mentions

Table 4
Semantic resources used to address bias in AI

Semantic high-level tasks	Resource	Reference
Bias assessment	Amazon KG	[29]
	Cellosaurus, DrugBank	[27]
	YAGO	[5]
	SentiWordNet	[18]
	prototype	[67]
	FrameNet	[57]
	Wikidata	[14]
	prototype	[38]
	WordNet	[17]
	proprietary	[31]
	SentiWordNet	[46]
	medical KG	[51]
Bias representation	Wikidata	[61]
	prototype	[47]
	MVSO	[43]
	IMAGACT	[34]
	DBpedia	[20]
	TIACRITIS	[80]
	CBOntology	[53]
	CODM	[22]
Bias mitigation	WordNet	[6]
	Wikidata	[54]
	prototype	[74]
	Wikidata	[3]
	DBpedia	[2,65]
	Freebase	[85]
	ConceptNet	[88]
	prototype	[24]
	ConceptNet	[21,39,41]
	DBpedia, WebChild	[21]
	ConceptNet, WordNet	[19]
	prototype	[76]

in the metadata and normalisation to map every instance to all its possible names. This analysis generated a new resource with unduplicated and normalised data that allows examination across study platforms. As a result, they identified that sex labels are inconsistently reported, with most samples lacking this information. More importantly, they report the existence of sex biases in drug data (e.g., female under-representation in studies of nervous system drugs). This study draws attention to the lack of study and the importance of including sex as a study variable in future analyses.

Another use of KGs relevant for this task is to assess disparities in the presentation of news reported by different sources, i.e., with different political leaning [5]. This *media bias* can affect groups or individuals who are part of the story or use these web search systems to build an opinion, since the news is written with the reporter's or media outlet's perception, which can be done, in some cases, partially or unfairly. As a result, bias compromises the reliability of the news source and may raise concerns closely related to the growth of misinformation, polarisation, or online hate. The structure of a KG could help to uncover such disparities in reporting between different media outlets, e.g., of specific stakeholders (politicians, political parties) advocating or opposing the same issue depending on which source the news appeared. The YAGO KG proved to help extract holders, opinions, and topics and store

them, allowing to compare topics and visualise biased news. Identifying potentially contradictory information is vital to raise awareness and encourage critical thinking because we, as web users, are exposed to a massive amount of information.

4.1.2. Bias affecting individuals

The assessment of polarised web search queries is also essential because users are prone to look for information that reinforces their existing beliefs [18]. Especially when it comes to controversial topics, the *confirmation bias* of web users often conveys views that can lead to a strong division of opinion, affecting individuals and society at large. The impact of bias in this type of user-generated content is closely related to the above mentioned media bias concerns. To prevent these issues, an approach to identifying the sentiment of queries could improve web search systems. This natural language processing (NLP) task (i.e. sentiment analysis) incorporates the support of the SentiWordNet lexical resource with a two-fold goal. First, it aims to improve the quality of results by including recommendations from less popular queries but with similar sentiments. More importantly, providing results from queries of opposite sentiment improves the diversity of opinions and shows the viability of query sentiment analysis to deal with the problem of bias and polarisation in web search.

Recently, a research project presented a similar approach to assess confirmation bias and other similar group phenomena that occur when analysing, visualising and disseminating information on the Internet (group polarisation [8] and the belief echo chamber [60]). The ontology is the primary data structure for modelling groups and individuals in a network structure [67] and is used as a tool to find similarities and anomalies in the profiles resulting from the data collected by the system. They present a web information system to evaluate such effects, integrating NLP methods into this data structure to identify themes and sentiments towards them. This architecture allows tracking individual and group responses to different events (e.g., COVID restrictions, vaccination) by simply adding new terms to the ontology (e.g., from specific vaccines such as Pfizer, Moderna or Astra Zeneca), inferring the sentiment towards them. These results are promising for integrating AI methods for natural language understanding (NLU) into distributed open ecosystems, as this is fundamental to understanding these phenomena in the broader societal domain.

4.1.3. Bias affecting AI systems

Bias can cause many problems affecting AI systems.

Description of approaches addressing the bias that leads to model overfitting.

Semantics are used to assess inconsistencies in the predictions of AI systems due to the use of a small, domain-specific corpus for training. *Model overfitting* can compromise the quality of AI systems and the extent to which they can fulfil their purpose. It is a significant challenge in the AI community due to the potential harms that may arise from using models that are black-boxes to us, especially when we do not understand why they make specific predictions. This survey found two use cases that focus on this problem. In the first example, sentences that entail the same action but are different may drift the model towards undesired behaviours, i.e., paying attention to the wrong words in the sentence [57]. The use of external knowledge from the FrameNet lexical database helped to reveal these biased predictions from the mismatch between the words in a sentence with the highest value in the attention layer of the model, and the action they should trigger, as captured in this semantic resource. This analysis revealed patterns with no theoretical basis but which the model systematically followed, i.e., recurrently giving more attention to words that were not relevant to trigger the action implied by the sentence. This paper also showed how this knowledge could be included as additional examples in the training data to make the model more consistent with the linguistic theory and help it generalise beyond the annotated examples of the training corpus.

Similar artefacts in datasets are evaluated in pre-trained masked linguistic models, which are increasingly used in factual knowledge bases to extract information from a query string [14]. The task consists of using queries such as "Steve Jobs was born in [MASK]", where *Steve Jobs* is the subject of the fact and *was born in* a prompt string for the relation "`place of birth`", to predict the object placed as [MASK]. However, their study demonstrates that many of the successes of *prompt-based* approaches are due to spurious correlations between similar prompts (Fig. 2). As a result, predictions on completely different datasets are similar, and this is because the dataset has been overfitted to specific prompts. The authors reveal that current *case-based* approaches that aim to improve performance by providing illustrative cases mainly succeed in providing a "type guidance". They reveal that performance is

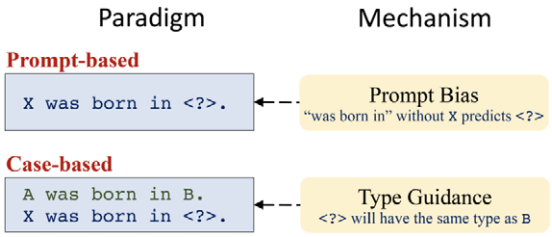

Fig. 2. Example of two dataset artifacts (i.e. *prompt bias, type guidance*) that can overfit pre-trained masked language models in factual knowledge extraction tasks [14].

enhanced primarily by recognising the object type in the illustrative cases. Therefore, models can effectively make analogies between entities of the same type but not predict facts based on their internal knowledge and the illustrative cases. This analysis is possible due to the use of the Wikidata taxonomy to infer the object type of each relation. It allows us to probe into the behaviours of these black-box models and better understand the critical factors underlying their task performance, which is crucial for building trust in the predictions of these systems in benchmarks and closed-world studies.

Recently, a new framework for automatic decision making in the medical domain was proposed to meet the requirements of explainability, robustness, and reduced bias in machine learning models [38]. Through a series of experiments to address three fundamental challenges in medical research, the authors reason that a multimodal, decentralised and explainable infrastructure is needed, where KG can play a crucial role. In this second use case, a series of arguments are presented as to why KGs can benefit future human-IA interfaces to be effective in this field (e.g., integrating the characteristics of different medical data modalities). It concludes with the basics of *counterfactual graphs*, which store the path from the feature to the changing class to enable the exploration of different counterfactual decision paths (bringing the "human-in-the-loop") and serves as a communication channel with black-box models. This work has a significant impact, as it provides knowledge-based constraints to regularise the training process of deep learning models and the possibility to contest them. This mechanism for opening the black box is critical, as future human-AI interfaces must enable medical experts to understand the causal pathways of automated decision-making systems.

Description of approaches addressing the bias that leads to human annotation errors.

Training corpora also have limitations due to errors in manual annotations that compromise the reliability of the corresponding AI systems. Subjectivity and errors in human annotations constitute a significant problem in developing benchmarks for AI systems, so the assessment of bias in annotation tasks is vital. In [17], background knowledge from the WordNet lexical database is used to support the annotation task where the context is missing, i.e., the set of neighbouring words that provide domain information. Notably, a comparison of the precision of two lexicographers in a context-agnostic scenario for a word sense disambiguation annotation task and using WordNet parameters to provide context revealed that annotations consistently shift towards the most frequent sense of a word in the absence of context. Even though this analysis used few semantic parameters (conceptual and semantic distance and belonging to the dominant concept), its binding machine versus human annotation study could help demonstrate the importance of context in human annotation tasks.

We found three other examples of assessment of *interpretation bias* in training data. One case study focused on assessing subjectivity in the interpretation of the analytical variables used to explain weather conditions in forecast texts, as these may vary due to humour, fatigue, or mood [31]. Using this as training data may compromise the truthfulness of the resulting weather prediction systems. They base their approach on the identification of numerical values and properties of different atmospheric variables in texts in order to be able to compare them with observational data. An ontology supports an information extraction model, as it can represent this domain knowledge. Specifically, they developed a proprietary ontology (*AEMIX*) using the Web Ontology Language (OWL) to extract the linguistic information of the critical events detected in the text and could reveal the inconsistencies of these texts with the objective information of the interpreted mathematical models.

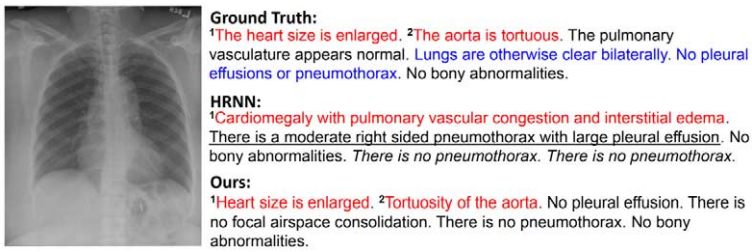

Fig. 3. Example of bias impact in image captioning for automatic radiology report generation when using human-generated data as ground truth for training automated decision-making systems [51].

Similarly, there can be bias when using user product reviews, blog posts and comments to support search engines, recommender systems, and market research applications, due to the subjectivity and ambiguity of this content [46]. As with the previous use case, this may compromise the quality and functionality of NLP systems, especially the degree to which they can detect mixed opinions. They propose a lexical induction approach because mapping subjectivity scores to opinion words in the text can detect review sentiment independently of individual language use. They use SentiWordNet to obtain these sentiment scores, which are used as additional input features for sentiment analysis. Despite the proposed method can only exploit several senses of each word (the overall score or the value of the first sense) without incorporating semantic relations, it shows the value of ontology-based approaches to avoid human biases arising from the use of machine-learned annotations.

Finally, we have found an excellent example of how inconsistencies in human-generated data used as ground truth to train automated decision-making systems can compromise the capability and effectiveness of these systems. The use of data-driven neural networks for automated radiology report generation is becoming a critical task in clinical practice [51]. In particular, image captioning approaches trained on medical images and their corresponding reports can significantly improve diagnostic radiology. However, the large volume of images that is a heavy workload for radiologists and, in some cases, lack of experience hinders the generation of these reports. The problem with using previous reports to train automation models is the variability and redundancy between the sentences used to describe the image, especially in describing the normal regions. For example, as shown in Fig. 3, the Blue text corresponds to the description of all normal image elements, while only the Red text indicates the abnormality. Since normal images are already overrepresented in the dataset, these deviations and repetitions aggravate the data imbalance and make the generation of sentences to describe normal regions more predominant. As a result, this bias in the data leads to errors where rare but significant abnormalities are not described (Underlined text) and repeated sentences describing the same normal region (*Italic text*). The use of prior medical domain knowledge captured in a KG showed an improvement in the reports generated, as seen in their quantitative and qualitative results (the text under "Our" in Fig. 3). In particular, the medical KG covering the most common abnormalities and findings can be used as an attention mechanism when exploring new input images. Its framework significantly improves abnormality detection, especially when the occurrence of normal reports dominates the entire dataset. We believe these enlightening examples can motivate future research, as similar data bias issues affect many AI applications.

4.2. Semantics to represent bias

There are some cases in which making the model's systematic preferences and possible biases transparent and expressing them in a human-understandable way may be decisive for providing possible directions of improvement [61]. Therefore, in the following section, we present examples of higher-level semantic tasks for bias representation.

4.2.1. Bias affecting specific groups of people

KGs could help to represent systematic preferences that are consistently applied across the examples used as training data [61]. This is the second example of *population bias*, as these preferences that influence the model, in the same way, are not individual features but domain categories representing specific groups. Mapping the influential input features to a KG (Wikidata) allowed them to be categorised and described with facts so that groups

corresponding to individual entities were captured as *counter-intuitive* rules. For example, that an Italian origin reduces the value of painters' works. Thus, this additional semantic reasoning capability revealed predictions based on undesirable input data features, such as race or gender, which can be critical for identifying the modification requirements that a model may need to mitigate biases.

There are three other examples of *population bias* where minority groups are disadvantaged by their lack of representation in the data. The first example focuses on the under-representation of minority cultures in music platforms, which leads to a dominance of commercial music and a lack of diversity [47]. Linked web data can be used to represent contextual features of music data (author biographical information or social connections between artists with similar singing patterns) and create a more relevant navigation space for the cultural background of music. Their prototype is based on a multimodal knowledge base, in which an Open Information Extraction system is used to extract contextual features of music data from the Linked Open Data (LOD). The combination of contextual features together with content features, extracted from audio recordings, can better contextualise the data and reveal non-trivial and deeper relationships between musical entities to lead to more meaningful music discovery and recommendations.

The use of ontologies can be advantageous in improving the representation of minority groups, as shown in the following two examples. In particular, being able to represent differences in the way of expressing and perceiving the affection conveyed by an image across languages can avoid the discrimination of generalist models that only fit a majority language [43]. Language is one of the main characteristics of a culture, so not paying attention to the context of each language can end up damaging entire ethnic groups. In this example, an ontology (Multilingual Visual Sentiment Ontology, *MVSO*) is constructed to represent a training dataset for visual sentiment analysis with a broader scope (including 12 different languages). Using NLP techniques, social media data in these languages and semantic resources (in particular, SentiWordNet and the SentiStrength ontology), they show that including this knowledge in the training datasets improves the degree to which image classification systems can predict sentiment for visual concepts in different languages. This work empirically demonstrates differences in model performance depending on the language used to express the sentiment used in training, as predictions from models trained in a specific language cannot generalise to image data collected in another language. Ensuring diversity in the training data is critical to avoid biased downstream applications to data from the predominant group.

On the other hand, the interpretation of which verbs trigger a given action varies from language to language [34]. The development of an ontology with videos to represent different actions enables identifying groups of verbs inclusive of different languages since videos have annotations in ten languages. The *IMAGACT* Ontology of Action is a video-based disambiguation framework that can help clustering algorithms not to be specific to a predominant language since they can thus rely on the multilingual lexical features of each action.

4.2.2. Bias affecting AI systems

We present a use case that uses a KG to improve the representation of users' interests beyond the items captured in the training data [20]. This example falls back to the *model overfitting* problem, which in this case can compromise the quality of recommender systems. Their framework incorporates a KG (DBpedia) to expand the user vector representation of a relational graph convolutional network used in the content-based RS. This structure encodes structural and relational information about the neighbouring nodes of the items already part of the training data to provide recommendations consistent with the users' needs. The propagation of relevant knowledge could enhance the performance of the recommender and dialogue systems.

4.2.3. Bias affecting individuals

As final examples, we present three case studies that use semantics to represent consistent and predicable errors that can compromise how data is used and analysed. This *psychological bias* can affect groups developing AI systems to support the search, interpretation, selection and visualisation of information needed to draw conclusions from large masses of data (Intelligence Activity, IA). The first example deals with its impact in the evaluation of evidence, in the search for hypotheses, and argumentation of scientific methods [80]. A domain ontology (*TIA-CRITIS*) is developed in a collaborative effort to represent all the reasoning steps, probabilistic assessments and assumptions of analysts in data-driven evidence analyses. Similarly, bias can affect planning, collection, processing and exploitation, analysis and production, dissemination and integration activities [53]. The *CBOntology* is an application ontology that captures the cognitive patterns known to affect these tasks in order to render them explicit

and support experts who may experience them. It covers more than 400 classes of such patterns extracted using string, semantic, logical, and topological matching similarities of existing ontologies. These two assistance tools aim to recognise known biases, advise the user to counter them and argue for the need to make biases explicit in AI systems and the experts who use them. The most recent surveyed paper related to this type of bias aims to support and reduce the psychological bias that occurs in decision making under risk and uncertainty [22]. Building on the Core Ontology on Decision Making (*CODM*), the authors extend this knowledge with the decision preferences that are bound to specific circumstances. For this, they use descriptive decision-making theory to extend the ontology with the concepts of intuitive decision-making so that choices made with deliberation or intuition can be explicitly represented to improve understanding of risk preferences and the situation in which they occur. All these works ultimately aim to develop decision support systems that help humans understand their own preferences to make better decisions.

4.3. Semantics to mitigate bias

The development of methods to mitigate bias is essential to prevent low-quality results that often impact communities and make them victims of policy injustice, affect their social perceptions, or disadvantage them in many AI application areas [6]. This section presents work examples that leverage semantic knowledge to counteract the possible adverse effects of biased learning.

4.3.1. Bias affecting specific groups of people

Semantic knowledge can be used to mitigate stereotypical perspectives of marginalised groups that are shown to be learned by automated decision-making systems [6]. This is another example of *population bias* because it reflects over-generalised beliefs about specific groups of people that can cause the model to shift towards incorrect predictions. In this example, hate speech detection systems are prone to be overly sensitive to the presence of specific demographic identity terms (gay, female, black) due to the large amount of hate content that exists against these communities. Their technique is based on replacing these bias-sensitive words with more abstract concepts, e.g., gay is a person, to prevent them from being incorrectly learned by the model as indicators of hate. They use WordNet's lexical relations to find suitable substitution candidates and demonstrate empirically that systematic deviations towards the hate class of these terms can be reduced without losing effectiveness in detecting hate speech. This initial data pre-processing bias mitigation technique reduces bias towards a closed list of words representing vulnerable groups.

4.3.2. Bias affecting individuals

KGs have shown advantages in mitigating biases due to humans' heuristic way in web searches. We present two use cases in which the structure of KGs can help counteract the *confirmation biases* that affect web users. On the one hand, one approach focuses on investigating the search environment to improve users' knowledge and attitudes on controversial topics (in particular, vaccination) [54]. The authors investigate including factual information extracted from Wikidata in a knowledge box in the search environment interface. It showed that users exposed to this information were significantly more informed, less sceptical about vaccination and more critical in discerning quality information after a simulated web search.

Similarly, the advantage of using a KG in the search interface is addressed in another study that aims to increase the efficiency, quality and user satisfaction with the information obtained after a web search [74]. To this end, they developed a KG-based interface prototype using the Open Information Extraction system to generate the entity-relationship-entity triplets of the text. Their qualitative study, based on a post-experiment evaluation, revealed that the KG interface helps to reduce the number of times required to view the source content during exploratory searches with respect to general hierarchical tree interfaces. These user-based studies serve to uncover important notions that shape the use of AI systems.

4.3.3. Bias affecting AI systems

Bias can cause many problems affecting AI systems.

Description of approaches addressing the bias that leads to human annotation errors.

Similar errors can affect the manual annotations often needed to train AI systems. The *scarcity* of such data compromises the development of systems to automate specific tasks. This use case assesses human annotators' errors due to a possible lack of knowledge to provide reliable annotations in extracting information from texts [3]. In particular, in coreference resolution tasks, annotators have to identify different mentions corresponding to the same entity. A reinforcement learning approach is proposed to address the lack of examples to train specific neural systems leveraging information from a knowledge base. Wikidata instances check the consistency of facts extracted from the text. Using this information to tune the model produces better results than other state-of-the-art methods and paves the way for assisting in the difficult task of obtaining human annotations needed in many AI systems.

Description of approaches addressing the bias that leads to data sparsity.

The rest of the use cases in this section deal with other limitations regarding how data is used to train AI systems. One of the problems of recommender applications is *data sparsity*, as less popular items are more challenging to deal with and may cause users to interact only with some of the most popular items [65]. This can pose a fairness problem because less popular items are under-represented, and so are the users that prefer to interact with less popular items. In this example, a framework is proposed to improve knowledge-based RS by including the specific semantic properties of the KG. Extracting each DBpedia property corresponding to user-item interactions allows computing similarity metrics between entities that consider each property's meaning. These property-specific interactions are included in the vectors that model the past interactions of each user to allow making more specific recommendations, e.g., movies that are related by the actors acting even if they do not deal with the same topic. This additional semantic knowledge of user-item interactions improved recommendations, especially on less popular data. Specifically, increasing the accuracy of recommended items after discarding the most obvious ones (serendipity) and the accuracy of unknown items that are part of the long tail of the catalogue (novelty).

Having many features in the recommendation motivates the proposal of an entropy-based method for obtaining only the meaningful historical data of each user. The sparse factorisation approach proposed in [2] facilitates the training process by exploiting a higher level of expressiveness in the feature embeddings of the items provided by a KG. Facts and knowledge extracted from DBpedia provide customised recommendation lists, filtering out items with low information gain. Their method allows incorporating the implicit information provided by the KG into the latent space of features, showing an improvement in the quality of results on three benchmark data compared to other state-of-the-art methods. More importantly, their experimental evaluation shows that this method improves item diversity, which is critical for measuring popularity bias mitigation.

Description of approaches addressing the bias that leads to missing data.

On the other hand, another limitation imposed by the platforms to collect information for recommendations is the small number of negative samples in the data, as most interactions are positive comments (clicks, purchases) [85]. This constitutes non-symmetric *missing data*, thus compromising population representation and potentially leading to biased analysis. A KG is proposed to provide informative negative signals to a collaborative filtering algorithm based on matrix factorisation. A negative sampler is constructed using reinforcement learning over Freebase to infer these signals from items related to positive interactions, assuming that these are more likely to be known to the user but were not chosen and, therefore, have a higher probability of being true negatives. Figure 4 is shown as an example. Given that a user (u_1) has watched two movies (i_1, i_2) with the same director (p_1) and genre (p_2), it is more likely that the user knows other movies (e.g., i_4) of the same director but different genre, for which the user has less interest. The reinforcement learning agent over a KG improved the top-K recommendation and preference ranking metrics of seven benchmark methods, which also used KGs, but only to leverage positive signals.

Description of approaches addressing the bias that leads to data imbalance.

Data imbalance can also affect the system's ability to infer new data from existing information. The following example aims to mitigate the bias caused by value propagation methods for sentiment analysis due to the imbalance between positive and negative seeds in the training data [88]. This imbalance causes new inferred values to drift towards the average value gradually. An additional step in the method is proposed to mitigate the bias on the basis that the propagation of values differs depending on the relationship between concepts, e.g., the relationship `isA` has a higher probability of concepts having the same sentiment value than other relations such as one concept `Desires`

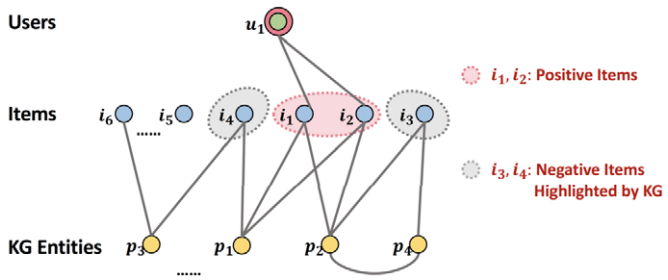

Fig. 4. Example of bias due non-symmetric missing data in a Recommender System (RS) that only collects positive samples [85].

another. Their method uses a sequential forward search over ConceptNet to select neighbouring concepts with the most relevant type of relationships to propagate new sentiment values. Next, the sentiment value concepts from a manually annotated sentiment dictionary (Affective Norms of English Words, ANEW) are used to align all inferred values with the mean and variance of the concepts that are in the dictionary, assuming that the difference between their inferred and original sentiment values is a shift that occurs due to the imbalance between the initial seeds.

Similarly, bias towards majority samples is a major challenge in current natural language generation (NLG) architectures. As a result, current neural approaches have difficulty generating coherent, grammatically correct text from structured data. The "divide-and-conquer" approach proposed in [24] is based on inducing a hierarchy from a corpus of unlabelled examples using a KG of entity and relation embeddings. The notion of similarity is used to show only the most relevant examples during training to avoid bias due to imbalances in the training data. Specifically, they apply this idea to two datasets containing linked data and textual descriptions (biography paragraphs with semantic mappings to Wikipedia infoboxes). Applying similarity of embedded inputs generates effective input-output pairs that consistently outperform competitive baseline approaches. Of particular interest in this study is partitioning the dataset according to the semantic and lexical similarity of the entries for training specialised models for each particular similarity group. This general idea can be transferred to other domains to address problems in data sparsity, such as image captioning or question answering applications.

Description of approaches addressing the bias that leads to model overfitting.

The following three examples address the problem of *model overfitting* by relying on the use of probabilistic models to generalise better cases that are not included in the training data. In image retrieval [41], relations to concepts in a KG can improve the reasoning power of the model in cases where examples of images with a given caption are not part of the dataset. In particular, the use of the ConceptNet commonsense base can be used to extend the search to images with related captions that are relevant, e.g., the concepts kitchen and restaurant can be informative of Chef. This approach can incorporate this rule-based knowledge source and enrich a language model widely used in multimedia-related tasks for NLP. Related concepts, i.e., relations that are relevant in visual space, are included in the object detection function of the model and show improvements in qualitative and quantitative results that are promising for the study of knowledge representation and computer vision.

Second, a similar approach has been applied in an image captioning framework to allow implicit image relationships to be captured in the caption that may be relevant to describe the image. For example, if the image shows a "woman standing with her luggage" next to a sign, then it makes sense to speculate that she is waiting for the bus [39]. The ConceptNet commonsense base helps discover these relationships, so a similar strategy is incorporated into the caption generator output to increase the likelihood of latent concepts related to specific objects in the image. This example leverages semantic knowledge to allow the system to generalise beyond the training examples.

Third, the incorporation of external knowledge can help mitigate the errors of systems that respond to questions related to an input image when dealing with answers that did not appear during their training phase or that are not contained within the image scene [21]. In real-world contexts, most techniques fail to address answers if they are not within the image content and require external knowledge. For example, Question 1 (Q1) in Fig. 5 requires such external knowledge, as responses such as "dog" (an entity that desires the frisbee) cannot be inferred from the image. A KG allows an understanding of the open-world scene beyond what is captured in the image. The framework

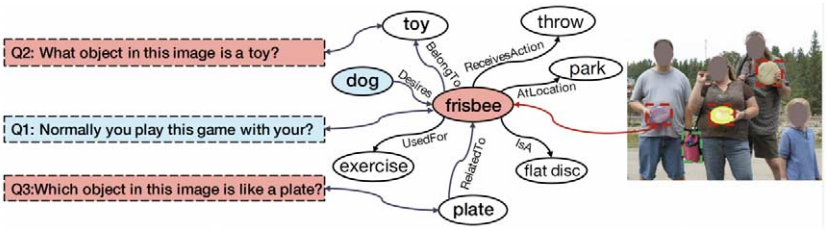

Fig. 5. Example of answer bias in a Visual Question Answering (VQA) system when dealing with concepts not seen during training or in the image scene [21].

incorporates prior knowledge to guide the alignment process between the feature embeddings of the image-question pair and the corresponding target response. Using a pre-existing subset of DBpedia, ConceptNet and WebChild, this knowledge component represents the possible set of answers and concepts to enrich the relations between them. Their quantitative results and their comparison with other current methods support the exploration of knowledge-based systems to overcome overfitting errors.

Description of approaches addressing the bias that leads to limited expressiveness when using the AI system.

Finally, two case studies have used semantics to mitigate problems in query formulations due to expressiveness limited to a small corpus leading to irrelevant and incorrect results. In image retrieval based on the semantic representation of scene graphs [19], WordNet and ConceptNet can be used to increase the precision of searches that include more complex concepts. For example, in cases where the system cannot infer that the entities "dog" and "cat" are relevant in the query "animals running on grass". Their approach introduces a set of rules to find images with fuzzy descriptions and infer the name of the concepts they express to help with more complex searches and enhance semantic and knowledge-based methods for image retrieval processing.

In order to reduce the number of irrelevant results of a web crawler to retrieve social media information, the development of a domain ontology is proposed [76]. Specifically, the Travel&Tour ontology is developed using the Protege-OWL editor to model the specific domain of travel and tourism. The aim is to enrich the content of social media data with the properties and relations of the ontology to improve context-specific searches and take advantage of the domain knowledge provided by this type of data. For example, a search for an expedition-type tourist destination in South America (`expedition+South+America`) may not return results because the extracted data does not explicitly mention those terms. However, there may be examples with mentions of related terms relevant to the search, e.g., Amazon river tours offered in the city of Brazil, even though there is no mention of the Amazon being in South America. Although these last two use cases presented only validate their results on a limited set of queries, they are initial works that favour data enrichment to alleviate the lack of expressiveness of the query methods.

5. Discussion

This section highlights the main findings whereby semantics can address bias in AI systems. In addition, it outlines the opportunities for contributions and challenges for further developments in ethical AI.

5.1. Major findings

We summarise some conclusions from the use case analysis following Table 5. We highlight the main issues and applications where semantics has helped and the tasks for which each semantic technology is most valuable. In particular, we focus our discussion along the following lines:

i) The use of semantics to address bias in AI is on the rise, particularly in approaches for mitigating, representing and assessing bias.
ii) The most researched application areas for biased AI systems using semantics are recommender and search systems and NLP applications.

Table 5

Full taxonomy of semantic tasks and bias in AI. Abbrv.: Recommender System (RS), Information Retrieval (IR), Information Extraction (IE), Natural Language Understanding (NLU), Natural Language Generation (NLG), Visual Question Answering (VQA), Text Classification (Text Clf.), Hate Speech detection (HS), Sentiment Analysis (SA), Word Sense Disambiguation (WSD), Music Search and Recommendation (Mus. S/R), Clustering (Clus.), Image Retrieval (Im. R), Image Captioning (Im. C), Image Sentiment Analysis (Im. SA), Intelligence Activity (IA), Content based Filtering (Cnt. F), Collaborative based Filtering (Col. F), Knowledge-based Recommender (K-based R), Scene Graph (SG), Search Engine (SE), Natural Language Processing (NLP), Machine Learning (ML), Computing (Comp.), Linked Open Data (LOD), Lexical Resource (Lexical R), Knowledge Graph (KG)

Type of bias	Bias location	Semantic high-level tasks	AI application	AI technology	SW Technology	Reference
Statistical	Functional	Assessment/Mitigation	RS	K-based R	KG	[29]
		Assessment	Medical Research	ML	Ontology	[27]
		Mitigation	RS	K-based R	KG	[2,65]
		Mitigation	RS	Col. F	KG	[85]
	Querying	Mitigation	Im. R	SG	KG	[19]
		Mitigation	IR	SE	Ontology	[76]
	Annotation	Assessment/Mitigation	NLU	NLP	Lexical R	[57]
		Assessment	IE	NLP	KG	[14]
		Assessment	Medical Research	ML	KG	[38]
		Representation	RS	Cnt. F	KG	[20]
		Mitigation	NLG	NLP	KG	[24]
		Mitigation	Im. R	NLP	KG	[41]
		Mitigation	Im. C	NLP	KG	[39]
		Mitigation	VQA	NLP	KG	[21]
	Aggregation	Mitigation	SA	NLP	KG	[88]
Cultural	External	Assessment	IR	SE	KG	[5]
		Representation	Text Clf.	NLP	KG	[61]
		Mitigation	HS	NLP	Lexical R	[6]
	Sampling	Representation	Mus. S/R	ML	LOD	[47]
		Representation	Im. SA	NLP	Ontology	[43]
		Representation	Clus.	ML	Ontology	[34]
Cognitive	External	Assessment	IR	SE	Lexical R	[18]
		Assessment	IR	SE	Ontology	[67]
	Functional	Mitigation	IR	SE	KG	[54,74]
	Annotation	Assessment	WSD	NLP	Lexical R	[17]
		Assessment	IE	NLP	Ontology	[31]
		Assessment	SA	NLP	Lexical R	[46]
		Assessment/Mitigation	Im. C	NLP	KG	[51]
		Mitigation	IE	NLP	KG	[3]
	Analysis	Representation	IA	Comp.	Ontology	[22,53,80]

iii) KGs are primarily used for bias mitigation, whereas ontologies are mainly used to represent bias and lexical resources and KGs to assess bias.

Incorporating formal knowledge representations into systems contributes to better generalisation, a fairer balance between bias and accuracy, and more robust methods. There is significant use of KGs to address these *technical challenges*, in particular, to mitigate bias in RS, but also in NLP tasks such as sentiment analysis, image retrieval, image captioning, natural language understanding, natural language generation and visual question answering. Major technical problems are sparsity, missing data, data imbalance, and overfitting due to small domain-specific training datasets.

The approaches to addressing sociological and psychological challenges have used a more varied range of semantic resources. On the one hand, minority ethnic groups are often less represented than the general population,

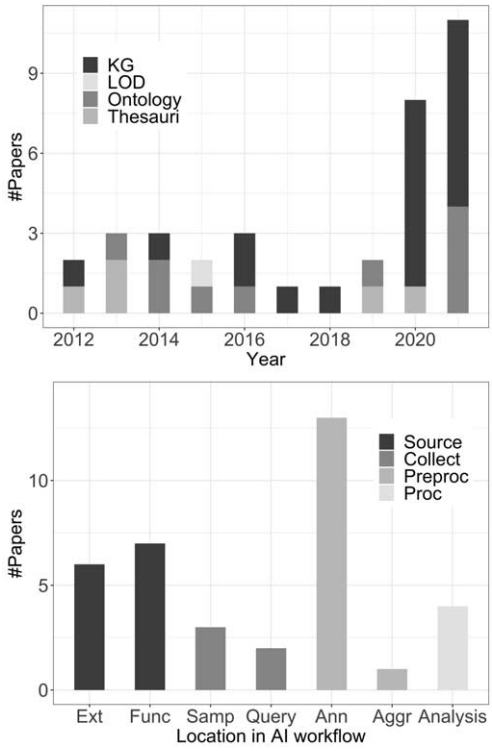

Fig. 6. Semantic web technologies used in this study (top-figure). Categories depending on the location in the AI workflow where bias originates (bottom-figure). Abbreviations: lOD (linked open data), KG (knowledge graph), source (bias at source), collect (bias at collection), preproc (data pre-processing), proc (data analysis).

which constitutes one of the main *sociological challenges* in AI systems. In this case, approaches generally focused on improving the representation of such groups (often reflected by their language) in different data modalities. In particular, we see how ontologies helped enhance diversity and inclusiveness in multimodal data (image and video), linked data in music data, and KGs and lexical resources in textual data.

On the other hand, the fact that many AI systems rely on human annotations to learn how to make their future predictions compromises, in many cases, their reliability and truthfulness. Given that human annotations are often liable to subjectivity, interpretation and lack of sufficient knowledge, it constitutes one of the major *psychological challenges* in AI. Previous works use ontologies and lexical resources to assess data annotation problems in NLP tasks (word sense disambiguation or information extraction from text), whereas KG appears promising for bias mitigation. Psychological challenges affect the creation of AI systems and how we interact with and use them. We have seen examples using lexical resources and KGs to assess and mitigate confirmation biases in web searches that affect how users interact with and process this content. Finally, ontologies can represent and make explicit the human psychology bias known to occur in computing activities to analyse and draw conclusions from data. We can conclude that semantics has contributed to addressing bias that can affect groups, individuals and AI systems, as it poses sociological, psychological and technical challenges.

5.2. Opportunities

We discuss the opportunities for semantics based on our major findings to elucidate the connection and future contribution to common AI bias methods. Specifically, we draw attention to the fact that:

i) KGs are likely to become increasingly dominant in AI bias research, given their wider scope and potential value in assessing and mitigating bias in various domains.

ii) Role of semantics in addressing bias related to the collection and annotation of data is likely to become a significant research direction in the near future.
iii) Semantic techniques will have a bigger role in enabling and enforcing fairness, explainability, and data pre-processing (including data augmentation, enrichment, and correction techniques).

From all the semantic resources that we have analysed in this work, KGs are particularly representative in the last two years (Top-Fig. 6). Thus, we expect their use to increase in the coming years. In particular, ConceptNet [19,39,41,88], DBpedia [20,65] and Wikidata [54,61] were mostly used in this study.

Bias mitigation approaches were the most representative of our study period. As seen in the lower part of Fig. 6, semantics can address bias at various stages of the AI workflow and especially the bias coming from the data annotation. The problem lies partly in human annotators' errors when processing information but crucially in the very limitations imposed by using annotated corpora to train AI systems. This is a general practice in AI but leads to systems with a capacity limited to the knowledge captured in a specific dataset.

Semantics shows great potential to contribute to several open lines of research on bias in AI. *Fairness metrics* are one of the most well-established approaches to avoiding bias and discrimination arising from the data or algorithms used. It is based on measures that evaluate the system's output concerning sensitive or protected attributes that should not affect the decision. Semantic knowledge seems promising for retrieving or approximating these characteristics, especially in cases where users have not disclosed this information [29]. Despite the variability and diversity of notation among existing fairness metrics [56], the richness of different perspectives around fairness will hopefully contribute to a better understanding of what fairness is and how to define it in AI systems.

We also emphasise the opportunities of semantics in future *human-AI* interfaces. We encountered several examples where the structure of KGs could enable addressing bias in the interaction with AI systems, e.g., when using web search engines [5,18,54,74]. Of particular interest is the vision of knowledge graphs as enablers of interactive and exploration-based explainability techniques [38], as integrating humans in the training, testing, and deployment phases of AI is necessary to bring these systems into real-world contexts.

We discuss the potential of semantics for developing data pre-processing techniques. We see an intersection of the area of knowledge representation with *data augmentation* techniques. These techniques rely on increasing samples to deal with unbalanced, unfair distributions or small data sets that may lead to discrimination of specific groups [44]. Semantics appears useful to re-sample from examples already existing in the data [85] or augment the dataset with new examples to make the model more consistent with the expected behaviours [3,57]. These approaches seem promising when the are disproportions between different classes or groups.

Furthermore, *data enrichment* techniques address bias by extending the features of instances that are already part of the dataset. In this case, there are several ways to incorporate semantic information into an enriched version of the features. Additional information about features (properties and relations) can be used as context to improve the generalisation of a specific dataset [47]. However, we must be aware of the noise that may result from the inclusion of uninformative features [71]. Another method is to extract patterns from the graph, e.g., to capture spurious model correlations that are based on sensitive information [61], or properties that enable mining less popular items in a RS [65]. Using KG to represent input features can improve generalisability for models trained with raw input features, as the graph structure gives a further analysis dimension. For example, for partitioning based on data imbalances using semantic similarity metrics [24]. Finally, probabilistic-based approaches to extend feature vector representations can generalise cases beyond the existing examples in the training dataset, increasing the likelihood of relevant entities given their semantic relationships to a data input. They offer advantages in multiple tasks (image captioning [39,51], image retrieval [41], visual question answering [21], and recommendation [2,20]). In summary, contextual enrichment, subgraph pattern mining and probabilistic-based approaches are promising research areas due to the increasing number of cases of individuals, groups, and AI systems still compromised by similar bias problems.

Finally, we discuss the opportunities of *data correction* techniques. Unlike the two previous approaches, these methods modify the data information to account for bias, maintaining the same number of samples and features. Semantic abstraction is a relevant concept in this respect, whereby the use of higher-level concepts of the information captured in the data may help generalising some dimensions that are not relevant to the task [75]. When data reflects bias and inequalities that the system can learn, such approaches seem worthwhile to retract and reduce the amount of information about specific groups that should not be retained by the model [6].

5.3. Open issues and challenges

This final part of the discussion highlights the challenges and issues we believe future research on AI bias will address.

There is great variability in the evaluation of AI bias-centred research works.

Generally, studies use two types of evaluation. User-based evaluation relies on user participation in the system through experimental or observational methods [34,54,74]. Besides, a common practice to evaluate the progress of AI systems is using baseline assessments, i.e., comparing approaches using benchmark datasets and specific algorithmic metrics.

The majority of works that address bias evaluate their approaches in downstream implementation. We found works using metrics generally used in recommendation and retrieval applications (ranking scores [20,21,41,85]) and NLP (textual similarity metrics [24,39,51], or general performance in multimodal [43] and text classification tasks [3,46,57,61,67,88]). It reveals the need to develop evaluation methods and metrics specific to bias, as an improvement in model performance generally does not reflect that the algorithm is not biased. The existence of a possible trade-off between overall performance and bias is an important topic of study in the bias literature [86]. However, only a few previous studies have considered formal definitions of fairness. For example, to ensure the quality and diversity of recommendations in individuals from disadvantaged groups [29] and underrepresented items [2], or to ensure that individuals from specific demographic groups are treated fairly by the system [6].

There is an ongoing debate about providing metrics that can be used to benchmark systems addressing bias. In many cases, evaluation frameworks account for demographic information about the individuals or groups affected by AI models. Still, it should take into account various forms of bias in existing models beyond the social categories that are considered as protected attributes by convention [10]. Moreover, these methods for measuring fairness can only reduce discrepancies about the characteristics captured in the data ("observed" space) [28]. While these may be relevant for prediction, they may not capture well all the characteristics that served as the basis for decision-making (the "construct" space), leading to the impossibility of a *fair* distribution.

Recommendation. The evaluation of forthcoming semantics-based methods to address bias requires a more critical evaluation that considers bias in the context of each particular application.

Secondly, semantic resources cannot be assumed to be free of bias.

Bias can be found in the data used to construct SW technologies.

The concept discussed in [25] of a *polyvocal* and *contextualised* SW draws attention to the fact that these knowledge sources often represent simplified views of the world, in which diverse perspectives may be underrepresented. In this light, the identification, representation and usage of different views or *voices* constitutes one of the main challenges in addressing that SW technologies often reflect the popularity or majority vote. Furthermore, web content is arguably increasingly centralised and asymmetric in terms of the distribution of knowledge and power. Thus, blockchain technologies present themselves as potential next-generation enablers of service exchange and content management [66]. A semantically enriched blockchain software ecosystem based on decentralised applications may be helpful to address bias due to less access of specific demographic groups to these technologies.

Previous works shed light on the *lack of representation* of specific groups in these technologies. An underrepresentation of less populated countries can occur in manually and semi-automatically created KGs such as Wikidata, as these consequently have a lower number of contributors [79]. Most worryingly, the correlation between coverage and population density is accurate in more developed countries but breaks for the large parts of Asia, Africa, and South America, where their content is drastically underrepresented. Such patterns were consistently found across the different language versions of DBpedia [42].

Demographic bias also propagates in automatic systems to generate KGs. Named entity recognition systems used for KG construction have shown a systematic exclusion in the detection of entities related to specific demographic categories like gender or ethnicity. For example, of black female names [59]. Gender disparities can occur in neural relation extraction systems when extracting specific links between entities (occupation [32]). As a result, bias

and prejudice against vulnerable demographic groups propagate into downstream applications. Such harmful associations of specific professions to particular gender, religion, ethnicity and nationality groups, such as men being more likely to be bankers and women to be homemakers, can be found in embeddings extracted from commonly used KGs (Wikidata, Freebase) [26]. Nevertheless, there are works to address how data representation disparities affect specific demographic groups. One example of a data augmentation method adds new samples to a KG to balance facts that regard specific sensitive attributes (gender differences in occupations) [69]. This approach effectively mitigates bias in the resulting embeddings from DBpedia and Wikidata. This example stresses the importance of bringing awareness and accounting for the possible bias arising from the application of semantic resources.

Besides the lack of representation, another bias assessed in particular semantic resources is mainly due to low coverage and noise. Content disparities may exist in different languages. For example, to address *limited coverage* in available general-purpose semantic resources, e.g., to only English, the authors in [82] propose a system to automatically extract lexical FrameNet units using Wikipedia pages in different languages as a reference. However, even when using the same language to model the same domain of knowledge, it is worth bearing in mind that there can be significant disparities between equivalent resources, e.g., as found in the synonym information of four lexical databases [81].

Statistical methods can serve to estimate the number of facts needed for relations to be representative of the real world [78]. Precisely, a method to calculate a lower bound of different relations could find that at least 46 million facts are missing to draw reasonable conclusions from DBpedia. Many of these missing entities may be due to the lack of type classes covered by DBpedia's ontology and used to automatically extract information from Wikipedia's infoboxes, which leads to only mapping a small subset to the graph [63]. Data descriptive methods appear as potential tools to assess these coverage problems, as shown in the analysis of missing data in specific languages in Wikidata [16].

On the other hand, *noise* is assessed in the annotation, generation, and evaluation of SW technologies. The suitability of the labels given to evaluate semantic-based systems objectively may be compromised, as seen in the differences between expert and crowdsource annotations of natural language summaries generated from a KG [83]. The achievement of link prediction (LP) approaches to automatically extend KGs may be obscured by the existence of inverse and symmetric relations in benchmark datasets. That is, achieving a good performance because certain relations in LP benchmarks tend to occur with others (the relation `born_in` with `located_in`), or have a default tail answer, as shown empirically in standard benchmarks extracted from Freebase, YAGO, and WordNet [72]. Consequently, the performance of LP methods diminishes tremendously in more realistic settings, e.g., by only removing inverse-duplicate relations from the benchmark dataset [1]. Probabilistic approaches can help return novel facts given redundant information. For instance, as shown in information extraction from tabular data to automatically complete KGs [48], where only entities well-covered were retained from the table.

Recommendation. Future semantics-based approaches to address bias in AI should ensure sufficient demographic representation of the people affected by the system and sufficient coverage of the application of use. Additionally, they should use such semantic information in realistic settings that account for noise, mainly due to redundant facts in the captured knowledge.

Consequently, it is imperative to increase transparency and explainability by publishing the source and currency of the data used to generate semantic resources [87], but equally the methods used to construct them. Especially in the enterprise, this information is crucial to ensure the integration of SW technologies in techniques to address bias in AI.

Bias can be found in the methods used to construct SW technologies.

Several factors introducing bias in the development of ontologies have been studied [45]. Specific philosophical views on whether an ontology should represent or interpret reality or its purpose constitute a bias arising from explicit choices. The same is true when capturing insights from competing scientific theories or when economic interests are at stake in deciding which domains deserve more attention. Other factors may propagate bias implicitly, such as specific levels of granularity, language, or underlying socio-cultural, political and religious motivations. There are examples of work that address these limitations, e.g., to define the scope of an ontology from the literature in a less biased way towards the selection of particular experts [36], or to compare the content coverage of the

ontology with a target domain [55]. These examples show the importance of raising awareness of the possible ethical implications of using these knowledge resources.

Similar problems affect the creation of general-purpose, like DBpedia, and domain-specific KGs, GeoNames. Increasingly, KGs are being integrated into search and recommendation systems to provide highly personalised content. The problems involved in creating such personalised KGs can be even more detrimental than in more conventional methods [33]. Specifically, this representation of users is at risk of being biased towards specific aspects depending on the data source used to collect information, e.g., behaviours on social networks are different from conversations in forums. In addition, timely events affect the type of information shared, e.g., in elections or times of pandemics. This bias can compromise user satisfaction and ultimately aggravate the echo chamber phenomenon. With the evolution and application of semantic technologies in new fields, it is important to be aware of new use issues.

Recommendation. The ethical implications of making particular decisions and selecting particular sources of information during the development of SW technologies require careful consideration when establishing the grounds for their application in techniques to address bias.

In conclusion, we draw on four main challenges posed by bias and prejudice in AI systems. First is the need to address the lack of data in unforeseen situations, i.e., shifting from controlled to open environments. We need to develop a world model that would enable AI for a general purpose and improve human-machine communication to use AI as a collaborative partner. Finally, we need to establish appropriate trade-offs between conflictive criteria to enable these systems to be applied in a broader range of applications. Integrating domain knowledge with data-driven machine learning models in a hybrid approach is key to addressing the identified challenges of *environment*, *purpose*, *collaboration*, and *governance* [40], so it is possible to develop ethically sensitive AI methods that work well in real-world applications. Therefore, while we need a critical analysis before applying semantics to address bias, the advances in new work and those seen in this article support research into knowledge-based reasoning techniques to overcome the pitfalls of current AI methods.

6. Conclusion

This survey shows the applicability of semantics to address bias in AI. From over a thousand initial search results, we follow a systematic approach and present the analysis of 34 use case studies that use formal knowledge representations (lexical resources, ontologies, knowledge graphs, or linked data) to assess, represent, or mitigate bias. We provide an ample understanding and categorisation of bias, discussing the harms associated with bias and the impact it can have on individuals, groups, and AI systems.

Our findings show that semantics has helped in many AI applications, including information retrieval, recommender systems, and numerous natural language processing tasks. Given the increasing use of semantics in recent years, particularly KGs, we conclude that semantics could primarily support fairness, explainability, and data preprocessing.

We identify further challenges in AI bias research for the SW and AI communities. These are primarily the need to develop more robust bias evaluation metrics beyond established sensitive information captured by dataset features that may not capture all the relevant information needed to build fair AI systems. We also discuss necessary considerations before applying SW technologies to address AI bias, such as underrepresentation of specific demographic groups, low application coverage, noise, and data source selection bias when building these technologies.

This paper positions the work of the SW community in the algorithmic bias context and analyses the intersection of both areas to assist future work in identifying and nurturing the benefits of these technologies, and using them responsibly. Bias in AI is an urgent issue because it compromises the applicability of automated systems in society, and semantics has enormous potential to help give meaning to the data they use.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie Actions (grant agreement number 860630) for the project "NoBIAS – Artificial Intelligence without Bias". This work reflects only the authors' views and the European Research Executive Agency (REA) is not responsible for any use that may be made of the information it contains.

References

[1] F. Akrami, L. Guo, W. Hu and C. Li, Re-evaluating embedding-based knowledge graph completion methods, in: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 1779–1782. doi:10.1145/3269206.3269266.

[2] V.W. Anelli, T. Di Noia, E. Di Sciascio, A. Ferrara and A.C.M. Mancino, Sparse feature factorization for recommender systems with knowledge graphs, in: *Fifteenth ACM Conference on Recommender Systems*, 2021, pp. 154–165. doi:10.1145/3460231.3474243.

[3] R. Aralikatte, H. Lent, A.V. Gonzalez, D. Herschcovich, C. Qiu, A. Sandholm, M. Ringaard and A. Søgaard, Rewarding coreference resolvers for being consistent with world knowledge, 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. 1229–1235, https://aclanthology.org/D19-1118. doi:10.18653/v1/D19-1118.

[4] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives, Dbpedia: A nucleus for a web of open data, in: *The Semantic Web*, Springer, 2007, pp. 722–735. doi:10.1007/978-3-540-76298-0_52.

[5] 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*, IEEE, 2012, pp. 878–881. doi:10.1109/ICDMW.2012.49.

[6] 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*, 2019, pp. 49–59. doi:10.1145/3308558.3313504.

[7] R. Baeza-Yates, Bias on the web, *Communications of the ACM* **61**(6) (2018), 54–61. doi:10.1145/3209581.

[8] C.A. Bail, L.P. Argyle, T.W. Brown, J.P. Bumpus, H. Chen, M.F. Hunzaker, J. Lee, M. Mann, F. Merhout and A. Volfovsky, Exposure to opposing views on social media can increase political polarization, *Proceedings of the National Academy of Sciences* **115**(37) (2018), 9216–9221. doi:10.1073/pnas.1804840115.

[9] S. Barocas and A.D. Selbst, Big data's disparate impact, *Calif. L. Rev.* **104** (2016), 671.

[10] 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, FAccT'21*, Association for Computing Machinery, New York, NY, USA, 2021, pp. 610–623. ISBN 9781450383097. doi:10.1145/3442188.3445922.

[11] C. Bizer, T. Heath and T. Berners-Lee, Linked data: The story so far, in: *Semantic Services, Interoperability and Web Applications: Emerging Concepts, IGI Global*, 2011, pp. 205–227. doi:10.4018/978-1-60960-593-3.ch008.

[12] S. Blodgett, S. Barocas, H. Daumé III and H. Wallach, Language (technology) is power: A critical survey of "bias", in: *NLP*, 2020, pp. 5454–5476. doi:10.18653/v1/2020.acl-main.485.

[13] 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.

[14] B. Cao, H. Lin, X. Han, L. Sun, L. Yan, M. Liao, T. Xue and J. Xu, Knowledgeable or Educated Guess? Revisiting Language Models as Knowledge Bases, 2021, arXiv preprint arXiv:2106.09231.

[15] 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), 1–14. doi:10.1186/s12859-018-2565-8.

[16] N. Chah and P. Andritsos, WikiMetaData studio: Dashboards from data profiling the languages, properties, and items of Wikidata, in: *Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) Co-Located with the 20th International Semantic Web Conference (ISWC 2021)*, 2021.

[17] A. Chatterjee, S. Joshi, P. Bhattacharyya, D. Kanojia and A.K. Meena, *A Study of the Sense Annotation Process: Man v/s Machine., in: GWC 2012 6th International Global Wordnet Conference*, 2012, p. 79.

[18] S. Chelaru, I.S. Altingovde, S. Siersdorfer and W. Nejdl, Analyzing, Detecting, and Exploiting Sentiment in Web Queries, *ACM Trans. Web* **8**(1) (2013). doi:10.1145/2535525.

[19] 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.

[20] 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, https://aclanthology.org/D19-1189. doi:10.18653/v1/D19-1189.

[21] Z. Chen, J. Chen, Y. Geng, J.Z. Pan, Z. Yuan and H. Chen, Zero-shot visual question answering using knowledge graph, in: *International Semantic Web Conference*, Springer, 2021, pp. 146–162.

[22] E. da Costa Ramosº, M.L.M. Campos, F. Baião and R. Guizzardiº, *Extending the Core Ontology on Decision Making According to Behavioral Economics*, 2021.

[23] E. Daga, M. d'Aquin, A. Adamou and S. Brown, The open university linked data–data, *open. ac. uk, Semantic Web* **7**(2) (2016), 183–191. doi:10.3233/SW-150182.
[24] N. Dethlefs, A. Schoene and H. Cuayáhuitl, A divide-and-conquer approach to neural natural language generation from structured data, *Neurocomputing* **433** (2021), 300–309. doi:10.1016/j.neucom.2020.12.083.
[25] M.v. Erp and V.d. Boer, A polyvocal and contextualised semantic web, in: *European Semantic Web Conference*, Springer, 2021, pp. 506–512. doi:10.1007/978-3-030-77385-4_30.
[26] J. Fisher, D. Palfrey, C. Christodoulopoulos and A. Mittal, Measuring social bias in knowledge graph embeddings, 2019, arXiv preprint arXiv:1912.02761.
[27] E. Flynn, A. Chang and R.B. Altman, Large-scale labeling and assessment of sex bias in publicly available expression data, *BMC bioinformatics* **22**(1) (2021), 1–23. doi:10.1186/s12859-020-03881-z.
[28] S.A. Friedler, C. Scheidegger and S. Venkatasubramanian, The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making, *Communications of the ACM* **64**(4) (2021), 136–143. doi:10.1145/3433949.
[29] Z. Fu, Y. Xian, R. Gao, J. Zhao, Q. Huang, Y. Ge, S. Xu, S. Geng, C. Shah, Y. Zhang et al., Fairness-aware explainable recommendation over knowledge graphs, in: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 69–78. doi:10.1145/3397271.3401051.
[30] F. Gandon, A survey of the first 20 years of research on semantic Web and linked data, *Revue des Sciences et Technologies de l'Information-Série ISI: Ingénierie des Systèmes d'information* (2018).
[31] A.L. Garrido, M.G. Buey, G. Muñoz and J.-L. Casado-Rubio, Information extraction on weather forecasts with semantic technologies, in: *International Conference on Applications of Natural Language to Information Systems*, Springer, 2016, pp. 140–151.
[32] A. Gaut, T. Sun, S. Tang, Y. Huang, J. Qian, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang et al., Towards understanding gender bias in relation extraction, 2019, arXiv preprint arXiv:1911.03642.
[33] E.J. Gerritse, F. Hasibi and A.P. de Vries, Bias in conversational search: The double-edged sword of the personalized knowledge graph, in: *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, 2020, pp. 133–136. doi:10.1145/3409256.3409834.
[34] L. Gregori, R. Varvara and A.A. Ravelli, Action type induction from multilingual lexical features, *Procesamiento del Lenguaje Natural* **63** (2019), 85–92.
[35] T. Gruber, Ontology, *Encyclopedia of database systems* **1** (2009), 1963–1965. doi:10.1007/978-0-387-39940-9_1318.
[36] M.K. Halawani, R. Forsyth and P. Lord, A literature based approach to define the scope of biomedical ontologies: A case study on a rehabilitation therapy ontology, 2017, arXiv preprint arXiv:1709.09450.
[37] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G.d. Melo, C. Gutierrez, S. Kirrane, J.E.L. Gayo, R. Navigli, S. Neumaier et al., Knowledge graphs, *Synthesis Lectures on Data, Semantics, and Knowledge* **12**(2) (2021), 1–257. doi:10.1007/978-3-031-01918-0.
[38] A. Holzinger, B. Malle, A. Saranti and B. Pfeifer, Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI, *Information Fusion* **71** (2021), 28–37. doi:10.1016/j.inffus.2021.01.008.
[39] 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*, 2020, pp. 535–544.
[40] A. Huizing, C. Veenman, M. Neerincx and J. Dijk, Hybrid AI: The way forward in AI by developing four dimensions, in: *International Workshop on the Foundations of Trustworthy AI Integrating Learning, Optimization and Reasoning*, Springer, 2020, pp. 71–76.
[41] 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, in: *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI'17*, AAAI Press, 2017, pp. 1283–1289. ISBN 9780999241103.
[42] K. Janowicz, B. Yan, B. Regalia, R. Zhu and G. Mai, Debiasing knowledge graphs: Why female presidents are not like female popes, in: *International Semantic Web Conference (P&D/Industry/BlueSky)*, 2018.
[43] 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*, 2015, pp. 159–168. doi:10.1145/2733373.2806246.
[44] 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.
[45] C.M. Keet, An exploration into cognitive bias in ontologies, in: *Proceedings of the Fifth Workshop on Cognition and OntologieS*, 2021.
[46] H.-J. Kim and M. Song, An ontology-based approach to sentiment classification of mixed opinions in online restaurant reviews, in: *International Conference on Social Informatics*, Springer, 2013, pp. 95–108. doi:10.1007/978-3-319-03260-3_9.
[47] G.K. Koduri, Culture-aware approaches to modeling and description of intonation using multimodal data, in: *International Conference on Knowledge Engineering and Knowledge Management*, Springer, 2014, pp. 209–217.
[48] B. Kruit, P. Boncz and J. Urbani, Extracting novel facts from tables for knowledge graph completion, in: *International Semantic Web Conference*, Springer, 2019, pp. 364–381.
[49] 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, 2020, pp. 410–416. doi:10.1109/BigComp48618.2020.00-39.
[50] B. Lepri, N. Oliver and A. Pentland, Ethical machines: The human-centric use of artificial intelligence, *IScience* **24**(3) (2021), 102249. doi:10.1016/j.isci.2021.102249.
[51] F. Liu, X. Wu, S. Ge, W. Fan and Y. Zou, Exploring and distilling posterior and prior knowledge for radiology report generation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13753–13762.

[52] H. Liu and P. Singh, ConceptNet—a practical commonsense reasoning tool-kit, *BT technology journal* **22**(4) (2004), 211–226. doi:10.1023/B:BTTJ.0000047600.45421.6d.
[53] 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)*, IEEE, 2014, pp. 237–243. doi:10.1109/CogSIMA.2014.6816616.
[54] R. Ludolph, A. Allam, P.J. Schulz et al., 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), e5430.
[55] R. Mac, An tsaoir, using spreading activation to evaluate and improve ontologies, in: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 2237–2248.
[56] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman and A. Galstyan, A survey on bias and fairness in machine learning, *ACM Computing Surveys (CSUR)* **54**(6) (2021), 1–35. doi:10.1145/3457607.
[57] 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: *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering (1)*, 2020.
[58] G.A. Miller, *WordNet: An Electronic Lexical Database*, MIT Press, 1998.
[59] S. Mishra, S. He and L. Belli, Assessing Demographic Bias in Named Entity Recognition, 2020, arXiv preprint arXiv:2008.03415.
[60] C.T. Nguyen, Echo chambers and epistemic bubbles, *Episteme* **17**(2) (2020), 141–161. doi:10.1017/epi.2018.32.
[61] 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*, 2020, pp. 1175–1184.
[62] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis et al., Bias in data-driven artificial intelligence systems—an introductory survey, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **10**(3) (2020), e1356.
[63] A. Nuzzolese, A. Gangemi, V. Presutti and P. Ciancarini, Type inference through the analysis of wikipedia links, *CEUR Workshop Proceedings* **937** (2012).
[64] A. Olteanu, C. Castillo, F. Diaz and E. Kıcıman, Social data: Biases, methodological pitfalls, and ethical boundaries, *Frontiers in Big Data* **2** (2019), 13. doi:10.3389/fdata.2019.00013.
[65] 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.
[66] T.G. Papaioannou, V. Stankovski, P. Kochovski, A. Simonet-Boulogne, C. Barelle, A. Ciaramella, M. Ciaramella and G.D. Stamoulis, A new blockchain ecosystem for trusted, traceable and transparent ontological knowledge management, in: *International Conference on the Economics of Grids, Clouds, Systems, and Services*, Springer, 2021, pp. 93–105. doi:10.1007/978-3-030-92916-9_8.
[67] M. Pavlícek, T. Filip and P. Sosí k, *ZREC Architecture for Textual Sentiment Analysis*, 2021.
[68] K. Petersen, S. Vakkalanka and L. Kuzniarz, Guidelines for conducting systematic mapping studies in software engineering: An update, *Information and software technology* **64** (2015), 1–18. doi:10.1016/j.infsof.2015.03.007.
[69] W. Radstok, M. Chekol, M. Schaefer et al., Are knowledge graph embedding models biased, or is it the data that they are trained on? in: *Wikidata Workshop 2021 Co-Located with the 20th International Semantic Web Conference (ISWC 2021)*, 2021.
[70] F. Richter and M. Sailer, *Basic Concepts of Lexical Resource Semantics*, 2003.
[71] S. Romero and K. Becker, Improving the classification of events in tweets using semantic enrichment, in: *Proceedings of the International Conference on Web Intelligence*, 2017, pp. 581–588. doi:10.1145/3106426.3106435.
[72] A. Rossi, D. Barbosa, D. Firmani, A. Matinata and P. Merialdo, Knowledge graph embedding for link prediction: A comparative analysis, *ACM Transactions on Knowledge Discovery from Data (TKDD)* **15**(2) (2021), 1–49.
[73] C. Roussey, F. Pinet, M.A. Kang and O. Corcho, *An Introduction to Ontologies and Ontology Engineering, in: Ontologies in Urban Development Projects*, Springer, 2011, pp. 9–38. doi:10.1007/978-0-85729-724-2_2.
[74] 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*, 2016, pp. 91–100. doi:10.1145/2854946.2854958.
[75] 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. doi:10.3233/SW-150188.
[76] E. Sediyono, C. Nivak et al., Measuring the performance of ontological based information retrieval from a social media, in: *2014 European Modelling Symposium*, IEEE, 2014, pp. 354–359. doi:10.1109/EMS.2014.15.
[77] T. Simonite, When it comes to gorillas, google photos remains blind, *Wired, January* **13** (2018).
[78] A. Soulet, A. Giacometti, B. Markhoff and F.M. Suchanek, Representativeness of knowledge bases with the generalized Benford's law, in: *International Semantic Web Conference*, Springer, 2018, pp. 374–390.
[79] D. Stepanova, M.H. Gad-Elrab and V.T. Ho, Rule induction and reasoning over knowledge graphs, in: *Reasoning Web International Summer School*, Springer, 2018, pp. 142–172.
[80] G. Tecuci, D. Schum, D. Marcu and M. Boicu, Recognizing and countering biases in intelligence analysis with TIACRITIS, in: *STIDS, Citeseer*, 2013, pp. 25–32.
[81] J. Teixeira, L. Sarmento and E. Oliveira, Comparing verb synonym resources for Portuguese, in: *International Conference on Computational Processing of the Portuguese Language*, Springer, 2010, pp. 100–109. doi:10.1007/978-3-642-12320-7_13.
[82] 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.

[83] P. Vougiouklis, E. Maddalena, J.S. Hare and E. Simperl, How biased is your NLG evaluation? (short paper), in: *Proceedings of the 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management (SAD 2018 and CrowdBias 2018) Co-Located the 6th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2018)*, Zürich, Switzerland, July 5, 2018, L. Aroyo, A. Dumitrache, P.K. Paritosh, A.J. Quinn, C. Welty, A. Checco, G. Demartini, U. Gadiraju and C. Sarasua, eds, CEUR Workshop Proceedings, Vol. 2276, CEUR-WS.org, 2018, pp. 72–77, http://ceur-ws.org/Vol-2276/paper8.pdf.

[84] D. Vrandečić and M. Krötzsch, Wikidata: A free collaborative knowledgebase, *Commun. ACM* **57**(10) (2014), 78–85. doi:10.1145/2629489.

[85] X. Wang, Y. Xu, X. He, Y. Cao, M. Wang and T.-S. Chua, Reinforced negative sampling over knowledge graph for recommendation, in: *Proceedings of the Web Conference 2020*, 2020, pp. 99–109. doi:10.1145/3366423.3380098.

[86] M. Wick, J.-B. Tristan et al., Unlocking fairness: a trade-off revisited, *Advances in neural information processing systems* **32** (2019).

[87] C.T. Wolf, From knowledge graphs to knowledge practices: On the need for transparency and explainability in enterprise knowledge graph applications, in: *Proceedings of the KG-BIAS Workshop 2020 at AKBC 2020*, 2020.

[88] 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.

[89] X. Zou, A survey on application of knowledge graph, *Journal of Physics: Conference Series* **1487** (2020), 012016.