AI & SOCIETY (2023) 38:747–761
https://doi.org/10.1007/s00146-022-01474-3

OPEN FORUM

Exposing implicit biases and stereotypes in human and artificial intelligence: state of the art and challenges with a focus on gender

Ludovica Marinucci[1] · Claudia Mazzuca[2] · Aldo Gangemi[1,3]

Received: 29 August 2021 / Accepted: 13 April 2022 / Published online: 26 May 2022

Abstract

Biases in cognition are ubiquitous. Social psychologists suggested biases and stereotypes serve a multifarious set of cognitive goals, while at the same time stressing their potential harmfulness. Recently, biases and stereotypes became the purview of heated debates in the machine learning community too. Researchers and developers are becoming increasingly aware of the fact that some biases, like gender and race biases, are entrenched in the algorithms some AI applications rely upon. Here, taking into account several existing approaches that address the problem of implicit biases and stereotypes, we propose that a strategy to cope with this phenomenon is to unmask those found in AI systems by understanding their cognitive dimension, rather than simply trying to correct algorithms. To this extent, we present a discussion bridging together findings from cognitive science and insights from machine learning that can be integrated in a state-of-the-art semantic network. Remarkably, this resource can be of assistance to scholars (e.g., cognitive and computer scientists) while at the same time contributing to refine AI regulations affecting social life. We show how only through a thorough understanding of the cognitive processes leading to biases, and through an interdisciplinary effort, we can make the best of AI technology.

Keywords Knowledge graph · Word embeddings · Implicit biases · Gender · Cognitive semantics

1 Introduction

In 2019, UNESCO published a report (West et al. 2019), borrowing its title I blush if I could from the response given by a feminine voice assistant to a human user exclaiming “Siri, you are a bi***!”. This report featured recommendations on actions to overcome gender gaps in digital skills, with a special examination of the impact of gender biases coded into some of the most prevalent AI applications. Indeed, in light of the explosive growth of digital voice assistants like Amazon’s Alexa or Apple’s Siri—often designed as feminine characters with subservient attitudes—recommendations concerning AI’s gender biases appear extremely urgent. Today, the answer of Apple’s AI assistant has been updated with a more neutral “I don’t know how to respond to that”, but despite the many efforts made so far, the rising problem of human-like biases in technological products still remains unsolved on both a practical and theoretical level. As the UNESCO report explains, these biases are rooted in gender imbalances in digital skills education, and thus in the gender imbalances of technical teams developing AI technologies for companies with significant gender disparities.

In the past few years, many researches unambiguously showed that gender biases, as well as racial biases, are found in Artificial Intelligence (AI). The AI-generated patterns, predictions, and recommended actions reflect the accuracy and reliability of the datasets used for training, as well as the assumptions and biases of the developers of the algorithms employed. Therefore, algorithms and devices have the potential of spreading and reinforcing harmful stereotypes. Such biases expose women, and especially women of color, at the risk of being left behind in economic, political, and social life. In fact, not only machine algorithms make movie

✉ Ludovica Marinucci
ludovica.marinucci@istc.cnr.it

Claudia Mazzuca
claudia.mazzuca@uniroma1.it

Aldo Gangemi
aldo.gangemi@cnr.it

[1] Institute of Cognitive Sciences and Technologies (ISTC), National Research Council (CNR), Rome, Italy

[2] Department Department of Dynamic, Clinical Psychology and Health, Sapienza University of Rome, Rome, Italy

[3] Present Address: Department of Classical and Italian Philology, University of Bologna, Bologna, Italy

recommendations, suggest products to buy, and perform automatic language translation, but they are also increasingly used in high-stakes decisions in health care systems (Obermeyer et al. 2019), bank loan applications (Mukerjee et al. 2002), hiring (Peng et al. 2019), and even in courts to assess the probability that a defendant recommits a crime. An example is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)[1] software, used in the United States (US) to decide whether to release or not an offender. An investigation into the software found a bias against African-Americans such that COMPAS was more likely to assign higher risk scores to African-American offenders than to Caucasians with the same profile (Dressel and Farid 2018). While AI poses significant threats to gender and racial equality, it is important to recognize that it also has the potential to make positive changes in our societies by showing us, and thus making us aware of, that the same automatic processes of our mind are embedded in the external objects we design.

The aim of our contribution is, therefore, to delineate a strategy allowing researchers to understand and identify biases in AI applications. This would subsequently allow the general public, such as stakeholders, but also people dealing with technology in everyday life, to become aware of the possible biases rooted in language, and consequently in machine learning applications trained on natural language. With this purpose, we first concentrate on biases and stereotypes from a cognitive science perspective (Sect. 2), with a special focus on gendered biases. Then, we turn to the description of how these biases are encoded in machine learning data (Sect. 3), expounding on the debate on whether these need to be corrected or simply acknowledged. In particular, we focus on a specific class of machine learning techniques, namely word embeddings, and we show how research dealing with this kind of techniques contributed to the overall awareness of gender biases. Finally,

we detail our proposal of an explainable human and artificial intelligence (Sect. 4) relying on the integration of both word embeddings and experimental cognitive data on the concept of *gender* in the state-of-the-art factual/linguistic resource Framester (Gangemi et al. 2016), making them available as a publicly dereferenceable and queryable knowledge base.

2 Biased minds: a top-down perspective on the debate among cognitive theories

Data coming from AI seem to be intrinsically biased. The reason underlying this issue lies in a very basic—but often underestimated—fact about text corpora on which these algorithms are trained: they are the result of human operations. And, as we are going to see, humans tend to be biased for specific cognitive reasons. From the first half of the twentieth century, the question of what human biases are animated scientific debates in cognitive science, and it still fuels academic discussions. Aware of the inevitable incompleteness of our analysis, below we present some of the most influential theories on cognitive biases.

2.1 Biases, heuristics, and stereotypes

In the late 1970s, Shiffrin and Schneider (1977) introduced a preliminary distinction between two forms of mental processing: conscious and controlled processing, on one side, and unconscious and automatic processing, on the other side. While conscious processing requires attention and motivation, which takes time to operate hence leading to a slow processing of information, automatic processing operates faster, outside of attention, and without executive control. According to Kahneman et al. (1982), only by knowing how these two parallel information processing, respectively, System 1 and 2, shape our judgments and decisions we can understand the profound effects of cognitive biases in the functioning of the human mind. In System 1, cognitive biases are linked to heuristics when individuals have to judge or make a decision under uncertainty. In fact, people's intuitive judgment deviates from the rules of logic or probability, challenging the idea that humans are rational beings (Tversky and Kahneman 1974). This deviation is explained by various kinds of heuristics and their related cognitive biases, such as the Adjustment heuristic leading to Anchoring effects, or the Representativeness heuristic leading to Base rate fallacy, Conjunction and Disjunction fallacy. Although subsequently criticized (Gigerenzer 1996), the heuristics-and-biases approach has changed the field of studies on cognitive biases by setting the main research questions, and permanently absorbing the concept of bias with that of error. In this framework, biases are, therefore, intended as problematic.

Other research programs have instead assimilated the notion of bias with that of stereotype. The latter is built on the two theoretical notions of associative networks in semantic memory and automatic activation. Concepts in semantic memory are assumed to be linked together in the form of associative networks, with associated concepts having stronger links, or being closer together, than unrelated

[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

concepts (Collins and Loftus 1975). A stereotypical association might be stored in semantic memory and automatically activated, hence producing an implicit stereotype effect (Devine 1989). Therefore, cognitive biases are involved in our way of classifying things: our expectations are automated so that the mere presence of a clue related to the category, that is a "category linked-cue", can activate a series of automatic associations without conscious awareness or intention of the individual (Devine and Sharp 2009). This automatic cognitive process plays a role not only in categorizing objects, but also in people, thereby giving rise to social stereotypes. Critically, automatic associations may also be found in individuals that do not share, or even repudiate, the content of such representations, as in the case of implicit race biases (Amodio and Devine 2006).

In this perspective, biases and stereotypes appear to be a pivotal feature of human categorization. "Stereotypes are fundamental to the ability to perceive, remember, plan, and act. Functionally, they may be regarded as mental helpers that operate in the form of heuristics or short-cuts." (Banaji 2002, 15,102). Likewise, Gladwell (2005) encourages us to give as much value to instantaneous impressions, fast as "the blink of an eye", as to months of rational analysis. Along these lines, cognitive biases are considered a type of heuristics, such as the "fast and frugal", that the mind needs to make better decisions in response to the stimuli of the complex external world (Gigerenzer et al. 1999). More recently, following Simon's (1955) idea of the "Bounded rationality" in decision-making with its "Satisficing" heuristic, Gigerenzer developed a systematic theory of heuristics called "Ecological Rationality". Under this account, the cognitive system relies on an adaptive toolbox, composed of biases people use for adaptive reasons, called "adaptive bias". By ignoring part of the available information due to his biased mind, his Homo Heuristicus can handle uncertainty more efficiently than an unbiased mind (Gigerenzer and Brighton 2009).

2.2 How to measure implicit biases in cognition?

Given the pervasiveness of biases and stereotypes in cognition, social psychologists started to search for suitable measures of implicit biases (e.g., Greenwald and Banaji 1995). With this purpose, Greenwald and collaborators (Greenwald et al. 1998) developed the Implicit Association Test (IAT), measuring reaction times to different combinations of word pairings with attributes presented over five sessions. In a standard IAT, participants are asked to quickly sort words into categories that are presented on the left or on the right of the screen, and that are represented by the target attributes and categories (e.g., flower, pleasant; insect, unpleasant). In the first session, participants are asked to sort pictures (or words) they are presented with based on word pairings (flower vs insect); in the second session, they sort them based on attributes (pleasant vs unpleasant); in the third session, word pairings and attributes are paired and participants are asked to sort pictures (or words) according to this parameter, while the last two sessions are counterbalancing sessions. When highly associated categories (e.g., flower and pleasant) share a response key, the reaction times are faster compared to the opposite condition (i.e., insect and pleasant sharing a response key). It is generally held that performance differences implicitly measure the degree of association of concepts and attributes.

The IAT task can be implemented with any word pair–attribute combination, but has been mainly used to examine a range of implicit stereotypes, such as racial and gender stereotypes (Greenwald et al. 2015). The results of these tests consistently reported the existence and endurance of negative stereotypical associations to certain social categories, such as people of color and women, hence suggesting these implicit biases are not the exception, but almost the rule in automatic information processing.

Stereotypical associations can even implicitly influence social judgements of people consciously seeking to avoid their use. It is still under debate whether and how cultural associations may be controllable and inhibited in decision-making, particularly when it comes to non-discriminatory judgments. For instance, Lai et al. (2016) examined nine intervention techniques aimed at reducing implicit racial bias, and showed that their effectiveness disappeared within a few days. So, implicit associations are malleable in the short term, but also firmly established in our minds.

On top of that, stereotypes may interfere with participants' cognitive performances, even when these are irrelevant for the task being performed. In a series of studies, white college participants who had previously taken a Race IAT were asked to interact with either a same-race or different-race peer confederate presented to them as the student manager of the test laboratory. Afterwards, they were asked to undertake an unrelated task, the Stroop color-naming task, generally used to measure executive control and cognitive depletion. The faster participants are at calling out the colors in which the words are printed, the higher their level of current executive control; the slower they are, the more cognitively depleted they are assumed to be. The findings show not only that the stress of interracial interaction undermined participants' subsequent executive control, but also that the greater was the relative ease with which they associated negative words with Black American racial categories in the IAT, the poorer was their Stroop performance after interracial interactions. This suggests that countering stereotypes is "cognitively costly" (Richeson and Shelton 2007, 317).

Whether biases and stereotypes are to be considered as "cognitive tools" or not, they can have disruptive consequences for the social life of specific social groups. To contrast this, it is essential to tackle the emergence and consolidation of cognitive mechanisms supporting the establishment in semantic memory of specific associations. Below we review some of the evidence documenting these processes with respect to a specific form of bias, that is gender bias.

2.3 Gendered biases: behavioral and linguistic evidence

Amongst several variables differentiating human beings, gender seems to be one of the most powerful and readily available cognitive tools. As early as 9 months, children can already distinguish female from male faces (Leinbach and Fagot 1993), and associate female voices to female faces (Poulin-Dubois et al. 1994). In addition, 6 years old children show high endorsement of essentialist theories of gender (Taylor et al.2009), suggesting the representation of gender as a category composed of two classes, each with its specific characteristics, might innately be essentialist. Additionally, it is claimed that gender is an evolutionarily salient perceptual feature; for instance, rapid gendered categorization based on physical cues occur as early as 150 ms after the stimulus onset, even when the task does not require directing attention to gender (Ito and Urland 2003).

It is, therefore, not surprising that gendered stereotypes and biases are so deeply rooted in our society. Physical and biological differences between women and men are often accompanied by the perception that these two classes are fundamentally different in terms of cognitive skills and behavioral attitudes. Research aimed at identifying psychological differences between women and men accumulated over the years (e.g., Ingalhalikar et al. 2014; Sax 2005), converging on the idea that these two discrete sexual categories are somehow different in their behavioral attitudes, cognitive capacities, and desires. While recent evidence coming from psychological meta analyses (e.g., Hyde 2005; Hyde et al. 2019) and neural findings (e.g., Joel et al. 2015) is nowadays challenging the idea of innate differences between women and men driving behavioral patterns, gendered stereotypes, and the biases they afford, remain partially unaffected.

One domain in which gender bias is especially evident is the one of careers. In this context, the conceptual association generally underlying patterns of gendered stereotypical thinking is the one that opposes women as primarily nurturing and affectionate, and men as competent and active (Bem 1974). This results in the belief that women are better suited for caring and family social roles, whereas men are thought to excel in social positions involving responsibilities and specific skills. For instance, results from over 600,000 online IAT tasks reported stereotypical associations between female terms and family or liberal arts, and male terms with science and career (Nosek et al. 2002). This kind of implicit association has a detrimental impact at the societal level, where male applicants are systematically favored to female applicants—even when there are no reliable differences in expertise. Moss-Racusin et al. (2012), for instance, compared the evaluations of applications for a laboratory manager position of faculty participants. Importantly, the application material was the same, but applicants were randomly assigned either feminine or masculine names. Their results show that male applicants were rated as significantly more competent and hireable than female applicants, despite their curricula being identical, and were assigned a higher starting salary compared to female applicants.

Not only gender stereotypes affect how the capabilities of women and men are perceived, but they also have an impact on how these two groups are evaluated and rewarded when it comes to their jobs. Along these lines, a meta-analysis of almost 100 studies targeting more than 200,000 employees from different industries (Joshi et al. 2015) revealed that performances of female workers tend to be judged significantly less favorably than those of male workers. On top of that, under the same objective circumstances, women are less likely to get promoted and reach prestigious positions, even in academic settings (Treviño et al. 2018).

It has been proposed that stereotypically desirable and undesirable traits for women and men relate to two basic semantic dimensions, namely potency and evaluation (Rudman et al. 2001). More recently, these two dimensions have been reinterpreted as opposing traits related to communality and warmth to those referred to agency and competence (Ellemers 2018); importantly, according to the Stereotype Content Model (SCM, see Fiscke et al. 2007) warmth and competence are also the two fundamental dimensions of social perception. Under this account, women would be mostly suited for caring social roles, whereas men would make the best of their jobs in agentic and powerful positions (for further discussions see Rippon 2019).

Language also encodes stereotypical associations. Cross-linguistic research showed that linguistic gender can impact the conceptual representation of even inanimate entities, such that in languages where grammatical gender is encoded speakers may conceptualize objects according to their grammatical gender (e.g., a table would be feminine in Spanish, La Mesa, and masculine in Italian, Il Tavolo, see Samuel et al. 2019 for a review). More to the point, not only linguistic structures affect conceptual representation, but stereotypical gendered associations are also carried by broader semantic associations. For instance, names that are grammatically masculine are more frequently rated as powerful and active than names that are grammatically feminine, which in turn are rated as prettier (e.g., Konishi

1993)—consistently with the opposition between warmth and competence already mentioned.

Within semantic domains, occupational nouns were found to be highly stereotypically gendered. Gender stereotypicality is understood as the belief that a certain social or occupational role is more likely to be occupied by one gender. This phenomenon has been measured through both implicit and explicit tasks, and was found to be consistent across cultures and languages (Misersky et al. 2014). In a semantic priming study composed of two experiments, Banaji and Hardin (1996) asked participants to judge whether pronouns they were presented with after role descriptors were feminine or masculine (Experiment 1), or to decide whether they were pronouns or not (Experiment 2). The role nouns used as primes were chosen based on census data on occupations, such that they reflected the actual skewness of the gender composition of occupations like nurse, secretary, doctor, mechanic. Participants were faster in responding after stimuli that were consistent with gendered expectations in both experiments—suggesting that stereotypical information about gender conveyed by language is encoded irrespectively of the task being performed. Online language processing also activates implicit stereotypical gender knowledge. Studies employing ERPs (Event Related Potentials, a measure used to determine the difficulty of processing certain stimuli) found that comprehending linguistic information consistent with stereotypical gender-expectations (e.g., feminine pronouns with the role descriptor nurse) is more fluent than comprehending inconsistent gendered information (e.g., masculine pronouns with nurse, see Misersky et al. 2019). Role nouns seem to be infused with gendered stereotypes even in the absence of grammatical cues denoting the gender of the referent (Gygax et al. 2008), so that English-speaking participants are more likely to associate *mathematician* with men than with women, even though the role-noun is not gendered in English (Misersky et al. 2014).

2.3.1 What is gender, and why is it so difficult to define?

The previous discussion outlined stereotypes and biases humans employ—sometimes unconsciously—to make sense of a gendered world. There are more than 7,500,000,000 people in the world, each with its own peculiarities and characteristics. It is, therefore, not surprising that our cognitive system employs short-cuts to broadly categorize people, and as discussed before, gender seems to be an evolutionary optimal candidate to serve as a categorizing cue. So, categorizing a person as either a woman or a man would tell us so much more than simply what their sexual organs are. It would give us hints into their behavioral and psychological attitudes, without delving further into their actual preferences. However, this account faces two main problems. On the one hand, it is clearly reductive, as it neglects individual variability in psychological and behavioral attitudes among people—while at the same time reinforcing gendered stereotypes proven false by recent evidence (e.g., Hyde 2005). On the other hand, it presupposes human beings are necessarily divided into two given classes, namely females and males, sharing some common traits based on biological differences. This assumption overlooks the variability of sexual configurations reported in biological studies (Fausto-Sterling 2012), and conflates sex into gender leaving non-conforming gender identities out of the picture.

These inconsistencies attest the tension between two opposing accounts on gender. Some research strands maintain that gender is rooted into biological sex differences between women and men, which further drive behavioral and cognitive differences (e.g., Ingalhalikar et al. 2014). This perspective is often referred to as 'biological/essential' theory (see Saguy et al. 2021). Under this account, gender is understood as an essential, objective, and natural category, stable across time and contexts, and composed of two fundamentally different classes, namely women and men. By contrast, social constructionist theories propose that gender is the result of sociocultural significations, and therefore that its boundaries are flexibly shaped by culture and society (West and Zimmerman 1987; Risman 2004; Butler 1990). In line with this perspective, gender differences are created by social factors and reinforced from infancy by differential treatment reserved to girls and boys. Fausto-Sterling et al. (2015), for example, showed that processes of gendered socialization occur as early as 3 months of age: mothers of daughters were more prone to take care of the appearance of their daughters, whereas mothers of sons engaged more frequently in rough motor activities. Anthropological and sociological findings also challenge the idea of gender binarism (i.e., the idea that there are only two classes of human beings) as a universal feature. In fact, social and cultural systems escaping the gender binary are widely documented across cultures and times (e.g., Hegarty et al. 2018).

More recently, some scholars proposed the label gender/sex (van Anders 2015; Hyde et al. 2019) to account for both biological and sociocultural components entrenched in the constitution of gendered and sexed identities. This proposal stresses the intertwinement between embodied (e.g., genitalia, hormones, bodily features) and sociocultural features (e.g., processes of socialization, cultural benchmarks), without necessarily laying on the essentialist or on the constructionist side of gender. So, while certainly less straightforward than some other accounts, hence more cognitively demanding, it seems particularly fitting to account for the complexity of gender. This view is also supported by research tapping into conceptual representations of gender. Indeed, in a recent study (Mazzuca et al. 2020a) gender was found to be conceptualized as a mixture of biological, perceptual features and sociocultural components. On top of that, its representation differed between "gender-normative" and "non-normative" individuals, suggesting personal and social experiences strongly influence the perception of gender.

As the previous discussion showed, human experience is heavily imbued with biases and stereotypes. Crucially though, human experience is a primary source of information, and information is what data are fed with. Machine learning algorithms trained on natural language data are, therefore, inevitably permeated with human biases and stereotypes. In the following section, we focus on recent discussions in the computer science literature with the aim of providing a further suggestion on the issue of gender-biased data.

3 Biased data: a bottom-up perspective on machine learning and natural language processing tools

With the commercialization and widespread use of AI systems and applications in our everyday lives, computer scientists in different subdomains such as machine learning, natural language processing, and deep learning are becoming increasingly aware of the biases that these applications can contain. A very detailed survey (Mehrabi et al. 2021) motivates researchers to tackle this issue by investigating different real-world applications that have shown unfair outcomes in the state-of-the-art methods. The survey provides a list of different sources and types of biases (such as the Representation bias, Sampling bias, Algorithmic bias, etc.), and examines how researchers have tried to address them. In the next, we pay special attention to what the authors call "historical biases" (Mehrabi et al. 2021, 4) in Word Embeddings (WE), and to how some interdisciplinary studies recently attempted to address this issue bridging the gap between results from cognitive psychology combined with those coming from WE.

3.1 Humans-in-the-loop: vicious or virtuous circle?

Recent advancements in the field of computer science revealed the multifaceted relation between cognitive biases and machine learning data—a relation that, as we are going to show, can lead to a vicious or virtuous circle.

On the one hand, close to machine learning applications, Kahneman and Tversky (1973) warned that cognitive biases can lead to violations of the Bayes theorem when people make fact-based predictions under uncertainty (see Sect. 2.1). Kliegr et al. (2021) discuss to what extent cognitive biases, as understood by Tversky and Kahneman (1974), may affect human understanding of interpretable machine learning models, in particular of logical rules discovered from data. Their review covered twenty cognitive biases, heuristics, and effects that can give rise to systematic errors when inductively learned rules are interpreted. For most biases and heuristics, psychologists have proposed "debiasing" measures that can be adopted by designers of machine learning algorithms and softwares. Application of empirical findings from cognitive science are also described to propose several methods that could be effective in suppressing these cognitive phenomena when machine learning models are interpreted. Finally, they suggest that future research should focus on empirical evaluation of the effects of cognitive biases in the machine learning domain.

On the other hand, since meaning is fundamental to many psychological processes, advances in the measurement of meaning, supported by WE (see Sect. 3.2), might also be of assistance to psychological sciences. Numerous researches validated WE as a means of representing the meanings contained in texts, demonstrating that WE retrieves known semantic and lexical relationships among words (e.g., Baroni et al. 2014). Because text represents an externalization of our semantic knowledge, psychologists are trying to adapt WE from computational linguistics to study the semantic organization of the human mind. In particular, WE has been used by different research strands, ranging from the study of decision-making (Bhatia 2017; Bhatia and Walasek 2019), language learning processes (Hollis 2017), brain imaging (Pereira et al. 2018; Zhang et al. 2020), to formal testing of psychological theories (van Loon and Freese 2019), and above all to model cognitive biases and stereotypes (Caliskan et al. 2017; Garg et al. 2018; Lewis and Lupyan 2020; Caliskan and Molly 2020).

In the following, we first focus on the debate among computer scientists on whether debiasing is necessary or not (Sect. 3.2); then we detail some of the most relevant interdisciplinary studies using WE to shed light on implicit biases and stereotypes encoded in natural language (Sect. 3.3).

3.2 The issue of debiasing word embeddings

Word Embedding (WE) represents a class of machine learning techniques used to uncover the semantic structure of text corpora. This recent and powerful machine learning technique is considered as a breakthrough in deep learning methods for its impressive performance on challenging natural language processing problems (Goldberg 2017). In WE, individual words are represented as real-valued vectors in a predefined vector space; each word is mapped to one vector and the vector values are learned in a way that resembles a neural network (Bengio et al. 2003). The learning process can be either joint with the neural network model on some tasks, such as document classification, or unsupervised, using document statistics.

State-of-the-art WE algorithms utilize neural networks to calculate the semantic relatedness of all words within a corpus on the basis of contextual interchangeability (Mikolov et al. 2017). Thus, words that occur in the same contexts are deemed more similar than words that occur in different contexts. In this way, WE can represent the relative meaning of all of the words within a language. The theoretical background of this approach is constituted by linguistic theories such as the distributional hypothesis (Harris 1954) according to which words occurring in similar contexts will have similar meanings. This notion is well expressed by Firth's notorious motto "You shall know a word by the company it keeps!" (Firth 1957, 11), implying that contextual information alone constitutes a viable representation of linguistic items. So, in WE the distributed representation is learned based on the usage of words, hence allowing words that are used in similar ways to have similar representations, and capturing their meaning.

Word2vec, a statistical method for efficiently learning a standalone WE from a text corpus developed by Google (Mikolov et al. 2013), has become a de facto standard for developing pre-trained WE. Originally conceived as an attempt to make the neural-network-based training of the embedding more efficient, it also featured the analysis of learned vectors, and the exploration of vector math on the representations of words. A well-known example shows that subtracting the "man-ness" from the word *King* and adding the "women-ness" produces the word Queen, capturing the analogy "king is to queen as man is to woman". Subsequently, the Global Vectors for Word Representation (GloVe) algorithm was developed by Stanford (Pennington et al. 2014) as an extension of the Word2vec method for efficiently learning word vectors. GloVe aims to marry two approaches: on the one hand, global statistics of matrix factorization techniques for classical vector space model representation, like Latent Semantic Analysis (LSA); while on the other hand, local context-based learning in Word2vec, which is more effective in capturing meanings by means of analogies. Among many types of intrinsic and extrinsic evaluations of these techniques (Bakarov 2018), solving word analogies has become one of the most popular benchmarks for WE, relying on the assumption that linear relations between word pairs such as king:man:: woman:queen are indicative of the quality of the embedding (Drozd et al. 2016).

The wide use of this intrinsic evaluation based on the completion of word vector analogies has brought unexpected results. The analogies showed that WE may carry biases mirroring those present in our societies, and thus encoded in our language. For example, in distributional semantics models (DSMs) like Word2vec and GloVe, the similarity of gender-neutral words like programmer with woman should not be lower than the similarity of programmer-man. However the arithmetic of these algorithms would solve the analogy reporting man:woman:: computer programmer:homemaker. This gave rise in 2016 to a heated and still ongoing debate within the machine learning community, aimed at identifying the best way to deal with this problem. The many attempts at reducing bias, either via post-processing (Bolukbasi et al. 2016) or directly in training (Zhao et al. 2019) have nevertheless left two research problems: (i) biases are still encoded implicitly in language, so that the effect of these attempts is mostly to hide them, rather than removing them. It has been claimed that, existing bias removal techniques are insufficient, and should not be trusted for providing gender-neutral modeling (Gonen and Goldberg 2019); more importantly, (ii) it is still under discussion whether we should aim at their removal or rather at transparency and awareness (Swinger et al. 2019), carrying out a fair analysis of human biases present in word embeddings, which cannot be addressed using analogy tasks (Nissim et al. 2020).

3.3 Leveraging word embeddings to expose implicit biases and stereotypes

Replicating a spectrum of known biases as measured by IATs (see Sect. 2.2), Caliskan et al. (2017) showed that training GloVe statistical machine learning model on a standard corpus of text from the World Wide Web, that is ordinary human language, can result in "human-like semantic biases" (Caliskan et al. 2017, 11). Furthermore, they developed the Word-Embedding Association Test (WEAT), a statistical test comparing word vectors for the same set of words used by the IAT. Contrarily to Bolukbasi et al. (2016), they rigorously demonstrated human-like biases in WE and hence suggested that WE not only track gender or ethnic stereotypes, but the whole spectrum of human biases entrenched in language. Indeed, they claimed that it would be impossible to use language significantly without incorporating biases, or as they put it: "we show that bias is meaning" (Caliskan et al. 2017, 12). So, text corpora contain imprints of implicit associations stored in our memory, which are often morally neutral, as in the case of insects and flowers.

While in some cases these associations can turn out to be discriminatory, sometimes they are merely veridical from an historical point of view. For example, Garg et al. (2018) used the temporal dynamics of WE to quantify historical trends and social changes in stereotypes and attitudes towards women and ethnic minorities in the twentieth and twenty-first centuries in the US. Integrating WE trained on 100 years of text data with the US Census, the results show that changes in the embedding track closely with demographic and occupation shifts over time. This approach is indeed a powerful intersection between machine learning and quantitative social science, which has been exploited to investigate the persistence of gender stereotypes too. Similarly, Lewis and Lupyan (2020) showed by means of IAT data from 39 countries assessing the association between women–family and men–career that countries with higher scores of implicit stereotypical gendered associations had also a lower percentage of women in STEM, as measured by UNESCO reports. Remarkably, using WE on two different corpora to retrieve words associated with females and males they also found a correlation between IAT scores and stereotypical gendered associations. This pattern held across 25 different countries and languages, such that participants whose dominant language showed stronger associations between women–family and men–career showed also stronger stereotypical associations in the IAT.

By reviewing these findings, Caliskan and Molly (2020) present evidence that word embeddings closely align with aspects of human cognition related to social reasoning—both in terms of implicit judgements and more objective social structural patterns and biases. However, they point out that while language statistics may have broad explanatory power in accounting for psychological constructs, the design of this approach is "correlational and thus unable to establish causality" (Caliskan and Molly 2020, 14). Therefore, the authors discuss two possible future directions for examining the extent to which language statistics play a causal role in shaping biases in human cognition: (i) a cross-linguistic generalization of the methods described to languages beyond English, and (ii) building causal models from observational data or experimental setting.

Along these lines, in the following section, we describe a further possible approach that arises in the direction of an explainable human and artificial intelligence, through the use of ontologies and knowledge representation (KR) integrating data from machine learning and cognitive psychology.

4 Awareness and transparency: the role of ontologies and knowledge representation in explainable human and artificial intelligence

A recent trend in artificial intelligence (AI) is trying to combine subsymbolic approaches (e.g., WE) with symbolic ones (e.g., Knowledge Graphs, KG)—as already proposed in a seminal discussion by Minsky (1991).

Along these lines, studies relating WE with WordNet are particularly relevant. WordNet (Fellbaum 1998) is a manually derived conceptual representation of word relations (e.g., synonymy, hyponymy, meronymy, etc.) based on psycholinguistic principles. WordNet has been extensively used in computational linguistics, but it can also be used as a knowledge graph (KG) (Fensel et al. 2020), and refined as a formal model of lexicon, thus gaining automated inferences and consistency checking from machine reasoners that use knowledge representation languages such as OWL (Allemang and Hendler 2011). In fact, WordNet has already been formalized in OWL/RDF (van Assem et al. 2006), and its structure has been reorganized under formal ontology principles in OntoWordNet (Gangemi et al. 2016), representing synsets (equivalence classes of word senses) and the other entities from WordNet as ontology elements (classes, properties, individuals, axioms), and linking them to the DOLCE foundational ontology[2] (Borgo et al. 2022). Additionally, recent studies have tried to automatically recreate WordNet's overall structure (Khodak et al. 2017) and substructure (Zhai et al. 2016) using information inferred from WE vectors. Specifically, Khodak et al. (2017) present a fully unsupervised method for the automated construction of wordnets based on a new word embedding-based method matching words to synsets, alongside the release of two large word-synset matching test sets for French and Russian. These experiments provide evidence that the relative position of WE vectors is reflective of semantic knowledge. WordNet was also one of the semantic resources used to test the AutoExtend system (Rothe and Schütze 2017), which formalizes it as a graph, where the objects of the resource are represented as nodes, and the edges describe relations between nodes. The nature of these relations can be either additive, when capturing the basic intuition of the offset calculus (Mikolov et al. 2013a), or based on similarity relations simply defining similar nodes. Based on these relations, the authors defined various constraints to select the set of embeddings that minimize the learning objective. For example, one constraint states that the embeddings of two synsets holding a similarity relation should be close.

While the potentiality of approaches combining WE and KG has been fairly explored—see also the Wembedder system of Wikidata KG (Nielsen 2017)—the use of these tools to address ethical issues in AI systems remains mostly unexploited. One notable exception is a recent study by Dancy and Saucier (2021), suggesting that "antiblackness" in AI requires more of an examination of the ontological space that provides a foundation for AI design, development, and deployment. To show an example of "antiblackness" they discuss results from auditing an existing open-source KG, called ConceptNet (Speer et al. 2017), that includes knowledge from several sources to connect terms with labeled, weighted edges. The ConceptNet API uses a system called "ConceptNet Numberbatch"[3] that combines data from several sources, such as ConceptNet 5, word2vec, GloVe, and

[2] http://www.ontologydesignpatterns.org/ont/dul/DUL.owl.

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

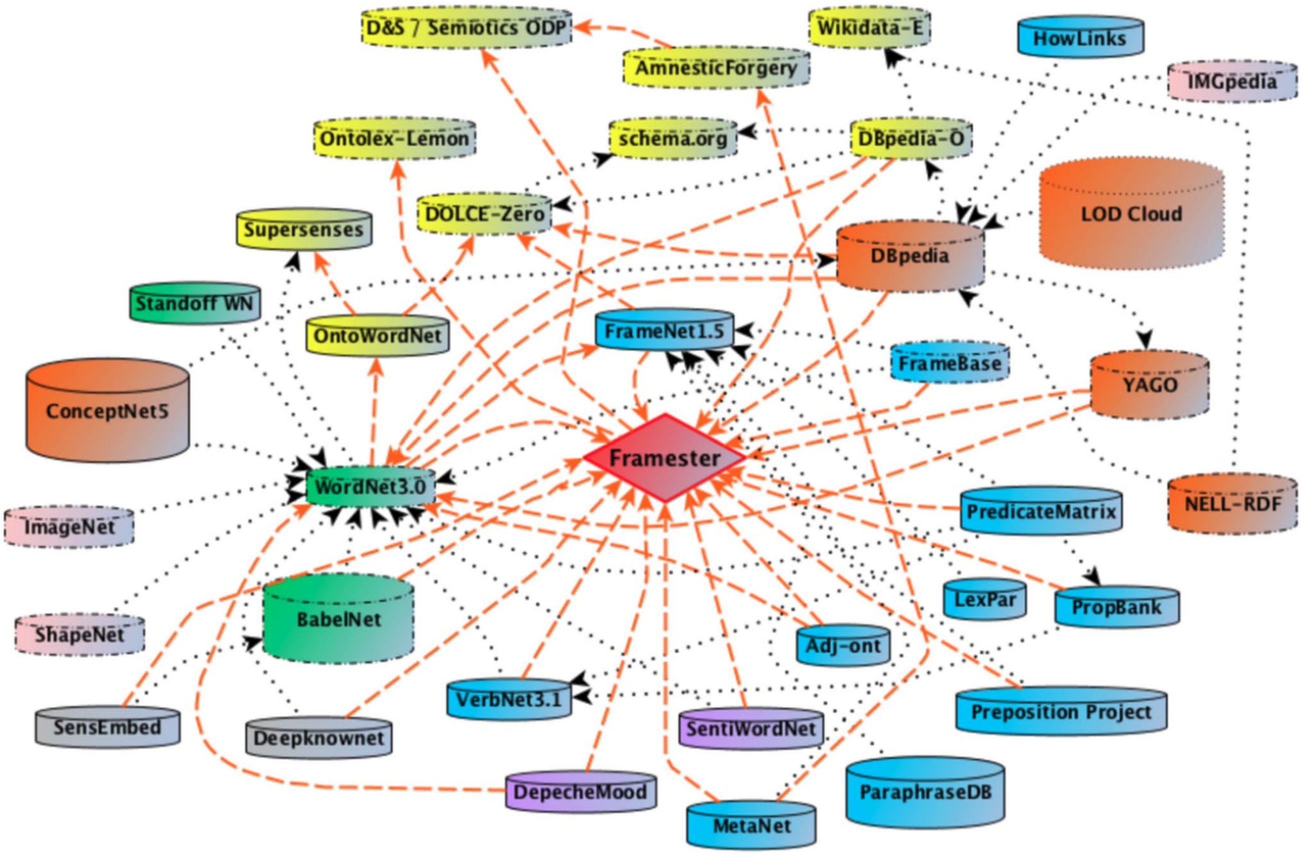

Fig. 1 Framester Cloud. Red represents Framester's main hub. Purple is for datasets for Sentiment Analysis. Orange arrows represent the Framester specific links, while Black arrows point to other existing links between the resources

OpenSubtitles 2016, by using a particular algorithm to calculate the relatedness for out-of-vocabulary terms. Despite the attempts to produce fairer relations between terms through "algorithmic de-biasing" (cf. Section 3.2), an exploration of semantic relatedness between the racialized terms black_man, white_man, black_woman, and white_woman, showed a pattern reflective of known historical relations and existing racial structures. For example, black_man retained a closer semantic relatedness to animalistic terms, while white_man retained human-related representations. It is notable—especially in a system that has been "de-biased"—the lack of semantic representation for black_woman, while white_woman remained a stand-in for woman, which shows the continued issue of intersectionality (Collins 2015), that is the issue of racialized, gendered intersections.

Furthermore, another notable exception among psychological investigations especially focusing on semantic memory is the study of Della Rosa et al. (2014), which used MultiWordNet[4] –a multilingual lexical database including an Italian version of WordNet that are aligned to external databases in other languages, such as Spanish, Portuguese, Hebrew, Romanian and Latin WordNets–to derive distinct types of abstract concepts. In particular, the authors of the experiment derived from MultiWordNet 5 distinct classes of abstract concepts, which are hyperonymy in WordNet, namely traits (e.g., weakness), actions (e.g., seduction), emotions (e.g., fear), social concepts (e.g., friendship), and cognitions (e.g., ideal). To overcome the fact that contents and the classification into different domains in WordNet are made a priori and do not take into account the distinction between abstract and concrete concepts, which is one of the core of the interpretative framework of the Word As social Tool (WAT) theory (Borghi and Binkofski 2014; Borghi et al. 2018), Della Rosa et al. (2014) additionally had a sample of participants rating concreteness, abstractness, and category membership.

Based on these findings, we emphasize that knowledge graphs using Linked Open Data (LOD) principles to provide easy access to structured data on the web can be used not only to contextualize and fix, if needed, AI systems making use of them, but also to understand human behavior and decision-making.

[4] https://multiwordnet.fbk.eu/english/home.php.

Table 1 Words resulting from the free-listing to the word "gender" with their frequency, and words resulting from word embeddings with their cosine value

Free-listing word	Percentage of participants producing the word (raw frequency)	WE from Wikipedia	Cosine
Identity	39 (30)	Sexuality	0.71
Sex	32 (25)	Grammatical	0.70
Female	26 (20)	Identity	0.69
Male	26 (20)	Sexual	0.68
Transgender	22 (17)	Orientation	0.67
Masculinity	21 (16)	Sex	0.64
Role	17 (13)	Role	0.61
Sexuality	17 (13)	Masculine	0.61
Equality	14 (11)	Neuter	0.59
Femininity	14 (11)	Plural	0.58

To achieve this goal, the integration between Natural Language Processing (NLP) and Semantic Web (SW) under the hat of "semantic technologies" requires a stable semantics allowing comparisons between tools or methods. In this sense, a wide coverage resource called Framester (Gangemi et al. 2016) can be particularly suitable for the goal of an explainable AI, as well as human intelligence, by creating an interoperable predicate space formalized according to frame semantics (Fillmore 1976), and ontological semiotics (Gangemi 2010). In fact, Framester is intended to work as a knowledge graph/linked data hub to connect lexical resources, NLP results, linked data, and ontologies. It uses the RDF versions of WordNet and FrameNet (Narayanan et al. 2000; Nuzzolese et al. 2011) at its core, and expands them transitively, by linking to lexical resources such as VerbNet (Kipper et al. 2000) and BabelNet (Navigli and Ponzetto 2012) as well as by reusing or linking ontological resources including OntoWordNet, DOLCE, Yago (Rebele et al. 2016), DBpedia (Auer et al. 2007), etc. Figure 1 displays all the resources integrated in Framester so far. Other new resources can be added in this dense interlinking by means of a homogeneous formalization for a direct and interoperable use of their data.

Therefore, in the following we propose an interdisciplinary perspective that aims at integrating both WE and free-associates with the concept of *gender* in Framester Linguistic Data Hub, making them available as a publicly dereferenceable and queryable knowledge base. To the best of our knowledge, there is to date no such resource combining insights from both machine learning and cognitive psychology that might help unravel the complexity of gender biases as exposed by the intertwinement between AI and cognitive processes.

4.1 Future directions: integrating word embeddings and free-listing data on the concept of gender

Our proposal of integrating free-associates and WE related to the concept of *gender* in an interoperable predicate space, based on foundational axioms (i.e., DOLCE ontology) applied to lexical knowledge (i.e., FrameNet, WordNet, etc.), has a two-folded objective. First, it aims at overcoming the lack of fairness and transparency in machine learning algorithms (Sect. 3); second, it is aimed at coping with the theoretical difficulty of defining what bias and stereotypes are (Sect. 2), with a focus on gender biases (Sect. 2.3). Finding this information in a structured formalization of a semantic network can be useful to a broad audience, ranging from developers looking for fairness in algorithms, to researchers in various disciplines willing to explore the resource for their studies, or even to legal and ethical experts for developing AI principles and regulations. Our assumption is that ontologies and knowledge representation can be key assets to enact hybrid systems, paving the way towards the creation of transparent and human-understandable intelligent systems.

As highlighted in Sect. 3.3, the information contained in WE models can be compared to self-report responses uncovering associations in natural language associations reflecting those present in our mind and society over time. Amongst numerous methods used to access semantic memory, semantic fluency tasks are consistently employed by researchers of diverse disciplines. Within this class of methods, free-listing tasks are frequently employed in neuropsychology to test the integrity of semantic knowledge in patients with brain injuries (e.g., Strauss et al. 2006), or in anthropology and linguistics to investigate the conceptual organization of specific cultural categories, such as "kinship", in different cultural and linguistic communities (Bernard 2006). In free-listing tasks, participants are presented with target words (e.g., "kinship"), and are asked to list as many exemplars related to the word as they can in a given timeframe. Free-listing tasks

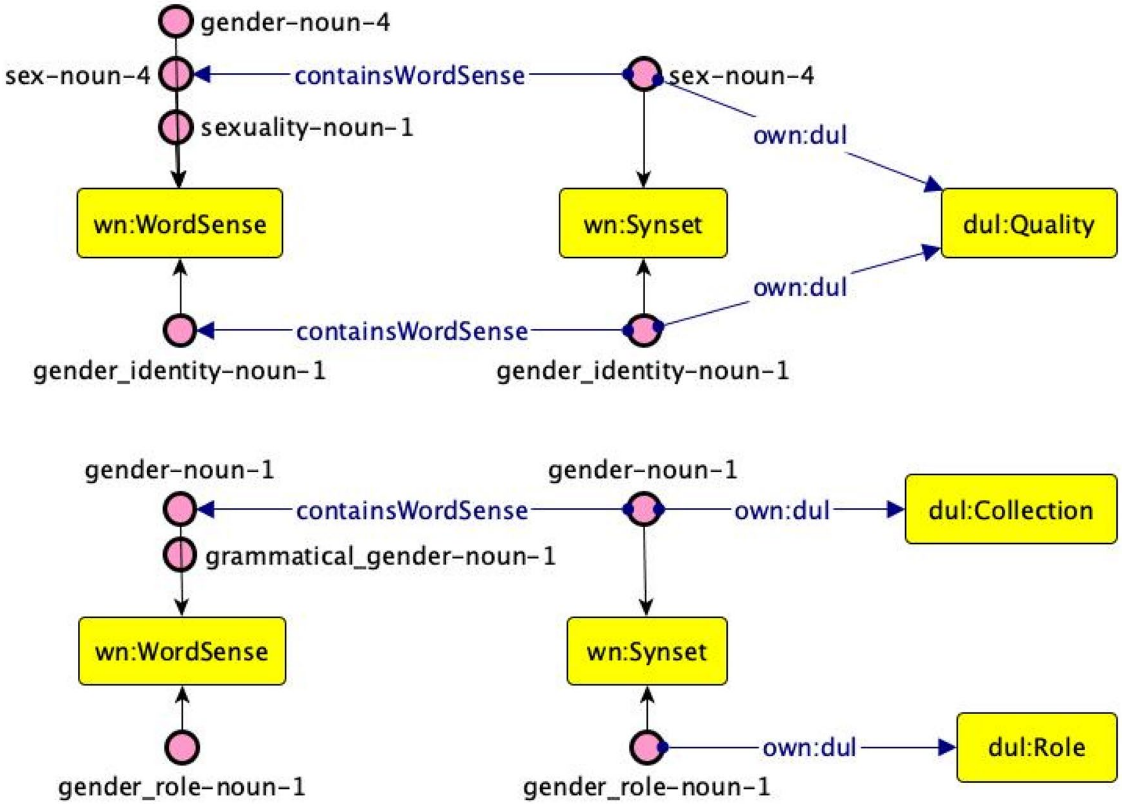

Fig. 2 Graphical representation in RDF/OWL of the main synsets and word senses of "gender" in WordNet. Each synset is also aligned to a DOLCE foundational category. Yellow and pink represent classes and instances respectively, while blue arrows point to object properties

have often been deployed to gain insights in the organization of conceptual knowledge by cognitive and developmental psychologists too, that by means of free-listing described the representation of several conceptual categories, such as animals (Crowe and Prescott 2003), landscape (van Putten et al. 2020), food (Hough and Ferraris 2010), and—more importantly for our purposes–gender (Mazzuca et al. 2020a).

Like other semantic fluency tasks, free-listing tasks are thought to provide an indirect measure of psychological proximity of concepts: the basic assumption underlying these kinds of tasks is that concepts that are mentioned earlier and more frequently are more psychologically salient for the target concept. In this sense, the rationale behind free-listing tasks can be somehow assimilated to that of WE, where words that are closer to the target word in a given corpus are thought to be more related to the target concept. Along these lines, we can compare WE for the target concept gender retrieved from Wikipedia English texts using GloVe (Mazzuca and Santarelli 2022) with free-listing data from an English-speaking sample of participants (Mazzuca et al. 2020b). Table 1 shows the top 10 words with higher cosine values for the WE data, and words listed at least by the 10% of participants in the free-listing task.

While Table 1 presents a comparison of results obtained with different methodologies sharing the underlying assumption that similar words go together (Firth 1957), our aim is to bring the analysis of these findings one step further. Specifically, we propose to match and link those words, without any kind of debiasing, to the related word senses in WordNet, as one of the core resources of Framester data hub. Therefore, we highlighted in boldface words listed in Table 1 (i.e., identity, sex, sexuality, grammatical, and role) that find a direct match with the WordNet word senses related to gender (i.e., gender-noun-4, sex-noun-4, sexuality-noun-1, gender_identity-noun-1, gender-noun-1, grammatical_gender-noun-1, gender_role-noun-1), as depicted in Fig. 2. This graphical representation in RDF/OWL (with Graffoo[5] notation) displays all the instances of the class wn:WordSense representing words with specific senses. As a set of synonyms, each word sense contributes to represent the concepts expressed by the istances of the class wn:Synset

[5] https://essepuntato.it/graffoo/.

(e.g., sex-noun-4). Additionally, in Framester each synset is aligned to a DOLCE foundational category. Among the DOLCE primitive classes, which have a relational axiomatization, dul:Quality represents the qualities specific to the entities to which they are inherent and, therefore, it depend on objects (dul:Object), abstracts (dul:Abstract), or events (dul:Event). In our example, both the instances sex-noun-4 and gender_identity-noun-1 are defined as dul:Quality, since they are inherent to the concept of gender either as a physical quality (sex-noun-4) or as a non-physical quality (gender-identity). Similarly, the entities in dul:Object class are divided into physical and non-physical. The latter includes socially constructed entities (dul:SocialObject) depending on physical objects to exist. Both the instances gender_role-noun-1 and grammatical_gender-noun-1 are included within this macro-category.

To summarize, linguistic associations retrieved either by prompting participants (free-listing task) or by means of machine learning techniques (WE) align with the semantic organization of WordNet synsets related to gender. In contrast with some accounts (see Sect. 2.3) these preliminary results relying on DOLCE classification suggest gender is mostly associated with non-physical and socially constructed concepts.

Additionally, the linking of WE and free-listing data to WordNet will also allow to find—by querying Framester SPARQL endpoint[6]—the frames they evoke in FrameNet, the sentiment and emotion scores in SentiWordNet (Baccianella et al. 2010) and DepecheMood (Araque et al. 2022), and any other data interlinked in the semantic network (cf. Figure 1) that may be interesting for a fine-grained analysis of the concept of gender.

The key examples in Fig. 2 shed light on the complexity of the semantic conceptualization of gender represented in a semantic network that exploits results from formal ontology investigations into the features that characterize conceptual distinctions, by applying foundational axioms to lexical knowledge. Here we proposed to enrich this structured knowledge with cognitive psychology and machine learning data, so as to gather evidence on conceptual representations and distinctions, less easily identifiable with one-way approaches. While the preliminary example we described deals specifically with the concept of gender, we suggest that embracing this approach might be a viable path to uncover and describe related representations of gendered biases and stereotypes. Importantly, this integrated resource responds to the need of accounting for conceptual associations entrenched in our language, that are subsequently embedded in text corpora—hence, exposing and possibly debunking biases and stereotypes found in AI applications.

5 Final remarks

The theoretical and technical discussions presented in our paper aimed at demonstrating the need for, and the fruitfulness of, interdisciplinary approaches and tools that can be put at service of an explainable human and artificial intelligence, and that further bridge the ever-narrowing gap between computer science and psycholinguistic studies. Taking into account several existing approaches that address the problem of biases and stereotypes, we also proposed the implementation of an integrated semantic resource that aims at integrating symbolic and subsymbolic knowledge. Auditing such a composited knowledge network provides an opportunity to probe implicit and hidden relations which are crucial to understand as long as they exist. This process becomes even more explicitly pivotal when these associations have a detrimental effect on the life of specific social groups.

All ethical debates (Mittelstadt et al. 2016) and frameworks to mitigate bias (Floridi et al. 2018) represent a superstructure of human decisions and responsibilities that the machine will have to learn. To be effective, this superstructure must rely on the awareness and consequent explainability of the deep cognitive structure and internal mechanisms of automatic and unconscious decision-making as a cognitive resource of all human beings. Only in this way AI applications can really get their positive value in our society, instead of being a source of dangerous unfairness. This synergic approach would thereby allow us to be able to learn from AI something about us—even if not necessarily pleasing—instead of only giving AI our data to learn.

Author contributions Ludovica Marinucci conceived the manuscript, and together with Claudia Mazzuca drafted it. Aldo Gangemi provided critical feedback and revised the manuscript.

Funding The authors received no fundings for this work.

Data availability Not applicable.

Materials availability Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest No potential competing interest is reported by the authors.

[6] http://etna.istc.cnr.it/framester2/sparql.

References

Allemang D, Hendler J (2011) Semantic web for the working ontologist: effective modeling in RDFS and OWL. Elsevier

Amodio DM, Devine PG (2006) Stereotyping and evaluation in implicit race bias: evidence for independent constructs and unique effects on behavior. J Pers Soc Psychol 91(4):652–661

Araque O, Gatti L, Staiano J, Guerini M (2022) Depechemood++: a bilingual emotion lexicon built through simple yet powerful techniques. IEEE Trans Affect Comput 13(1):496–507

Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z (2007) DBpedia: a nucleus for a web of open data. In: Aberer K et al (eds) The Semantic Web. ISWC 2007, ASWC 2007. Lecture notes in computer science, vol 4825. Springer, Heidelberg, pp 722–735

Baccianella S, Esuli A, Sebastiani F (2010) Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA)

Bakarov A (2018) A survey of word embeddings evaluation methods. arXiv preprint arXiv:1801.09536

Banaji MR (2002) Stereotypes, social psychology of. International encyclopedia of the social and behavioral sciences, 15100–15104

Banaji MR, Hardin CD (1996) Automatic stereotyping. Psychol Sci 7(3):136–141

Baroni M, Dinu G, Kruszewski G (2014) Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In: K. Toutanova and H. Wu (Eds.), Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 1, 238–247

Bem SL (1974) The measurement of psychological androgyny. J Consult Clin Psychol 42(2):155–162

Bengio Y, Ducharme R, Vincent P, Janvin C (2003) A neural probabilistic language model. J Mach Learn Res 3:1137–1155

Bhatia S (2017) Associative judgment and vector space semantics. Psychol Rev 124:1–20

Bhatia S, Walasek L (2019) Association and response accuracy in the wild. Mem Cognit 47:292–298

Bernard HR (2006) Research methods in anthropology: qualitative and quantitative approaches. AltaMira Press, Lanham

Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT (2016) Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Adv Neural Inf Process Syst 29:4349–4357

Borghi AM, Binkofski F (2014) Words as social tools: an embodied view on abstract concepts, vol 2. Springer, New York

Borghi AM, Barca L, Binkofski F, Tummolini L (2018) Varieties of abstract concepts: development, use and representation in the brain. Phil Trans R Soc B. https://doi.org/10.1098/rstb.2017.0121

Borgo S et al (2022) DOLCE: a descriptive ontology for linguistic and cognitive engineering. Appl Ontol 17(1):45–69. https://doi.org/10.3233/AO-210259

Butler J (1990) Gender trouble: feminism and the subversion of identity. Routledge, New York

Caliskan A, Molly L (2020) Social biases in word embeddings and their relation to human cognition. PsyArXiv Preprint. https://doi.org/10.31234/osf.io/d84kg

Caliskan A, Bryson JJ, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. Science 356:183–186

Collins PH (2015) Intersectionality's definitional dilemmas. Ann Rev Sociol 41(1):1–20

Collins AM, Loftus EF (1975) A spreading-activation theory of semantic processing. Psychol Rev 82(6):407–428

Crowe SJ, Prescott TJ (2003) Continuity and change in the development of category structure: insights from the semantic fluency task. Int J of Behav Dev 27(5):467–479

Dancy CL, Saucier PK (2021) AI and blackness: towards moving beyond bias and representation. IEEE Trans Technol Soc. https://doi.org/10.1109/TTS.2021.3125998

Della Rosa PA, Catricalà E, De Battisti S, Vinson D, Vigliocco G, Cappa SF (2014) How to assess abstract conceptual knowledge: construction, standardization and validation of a new battery of semantic memory tests. Funct Neurol 29(1):47–55

Devine PG (1989) Stereotypes and prejudice: their automatic and controlled components. J Pers Soc Psychol 56(1):5–18

Devine PG, Sharp LB (2009) Automaticity and control in stereotyping and prejudice. Handbook of prejudice stereotyping, and discrimination. American Psychological Association

Drozd A, Gladkova A, Matsuoka S (2016) Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In: Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers, 3519–3530

Dressel J, Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. Sci Adv. https://doi.org/10.1126/sciadv.aao5580

Ellemers N (2018) Gender stereotypes. Annu Rev Psychol 69:275–298

Fausto-Sterling A (2012) Sex/gender: biology in a social world. Routledge, New York

Fausto-Sterling A, Crews D, Sung J, García-Coll C, Seifer R (2015) Multimodal sex-related differences in infant and in infant-directed maternal behaviors during months three through twelve of development. Dev Psychol 51(10):1351

Fellbaum C (1998) WordNet: an electronic lexical database. Bradford books. The MIT Press

Fensel D et al (2020) Knowledge graphs. Springer International Publishing, Switzerland

Fillmore CJ (1976) Frame semantics and the nature of language. Ann N Y Acad Sci 280(1):20–32

Firth, J. R. (1957). A synopsis of linguistic theory 1930–55. Reprinted in: FR Palmer (ed.). (1968). Selected Papers of JR Firth 1952, 59.

Fiske ST, Cuddy AJ, Glick P (2007) Universal dimensions of social cognition: warmth and competence. Trends Cogn Sci 11(2):77–83

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V et al (2018) AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind Mach 28(4):689–707

Gangemi A (2010) What's in a schema? A formal metamodel for ECG and FrameNet. Ontology and the Lexicon: a natural language processing perspective. Cambridge University Press, Cambridge, pp 144–182

Gangemi A, Alam M, Asprino L, Presutti V, Recupero DR (2016) Framester: a wide coverage linguistic linked data hub. European knowledge acquisition workshop. Springer, Cham, pp 239–254

Garg N, Schiebinger L, Jurafsky D, Zou J (2018) Word embeddings quantify 100 years of gender and ethnic stereotypes. Proc Natl Acad Sci 115:3635–3644

Gigerenzer G (1996) On narrow norms and vague heuristics: a reply to Kahneman and Tversky. Psychol Rev 103(3):592–596

Gigerenzer G, Brighton H (2009) Homo Heuristicus: why biased minds make better inferences. Top Cogn Sci 1(1):107–143

Gigerenzer G, Todd PM, the ABC Group (1999) Simple Heuristics that make us smart. Oxford University Press, Oxford

Gladwell M (2005) Blink: the power of thinking without thinking. Little Brown, New York

Goldberg Y (2017) Neural network methods for natural language processing. Synth Lect Hum Lang Technol 10(1):1–309

Gonen H, Goldberg Y (2019) Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 609–614

Greenwald AG, Banaji MR (1995) Implicit social cognition: attitudes, self-esteem, and stereotypes. Psychol Rev 102(1):4–27

Greenwald AG, McGhee DE, Schwartz JL (1998) Measuring individual differences in implicit cognition: the implicit association test. J Pers Soc Psychol 74(6):1464

Greenwald AG, Banaji MR, Nosek BA (2015) Statistically small effects of the Implicit Association Test can have societally large effects. J Pers Soc Psychol 108(4):553–561

Gygax P, Gabriel U, Sarrasin O, Oakhill J, Garnham A (2008) Generically intended, but specifically interpreted: When beauticians, musicians, and mechanics are all men. Lang Cognit Process 23(3):464–485

Harris ZS (1954) Distributional structure. Word 10(2–3):146–162

Hollis G (2017) Estimating the average need of semantic knowledge from distributional semantic models. Mem Cogn 45:1350–1370

Hegarty P, Ansara YG, Barker MJ (2018) Nonbinary gender identities. In: Dess NK, Marecek J, Bell LC (eds) Gender, sex, and sexualities: psychological perspectives. Oxford University Press, Oxford, pp 53–76

Hough G, Ferraris D (2010) Free listing: a method to gain initial insight of a food category. Food Qual Prefer 21(3):295–301

Hyde JS (2005) The gender similarities hypothesis. Am Psychol 60(6):581–592

Hyde JS, Bigler RS, Joel D, Tate CC, van Anders SM (2019) The future of sex and gender in psychology: five challenges to the gender binary. Am Psychol 74(2):171–193

Ingalhalikar M, Smith A, Parker D, Satterthwaite TD, Elliott MA, Ruparel K et al (2014) Sex differences in the structural connectome of the human brain. Proc Natl Acad Sci 111(2):823–828

Ito TA, Urland GR (2003) Race and gender on the brain: electrocortical measures of attention to the race and gender of multiply categorizable individuals. J Pers Soc Psychol 85(4):616–626

Joel D, Berman Z, Tavor I, Wexler N, Gaber O, Stein Y et al (2015) Sex beyond the genitalia: the human brain mosaic. Proc Natl Acad Sci 112(50):15468–15473

Joshi A, Son J, Roh H (2015) When can women close the gap? A meta-analytic test of sex differences in performance and rewards. Acad Manag J 58(5):1516–1545

Kahneman D, Tversky A (1973) On the psychology of prediction. Psychol Rev 80:237–251

Kahneman D, Slovic B, Tversky A (1982) Judgment under uncertainty: heuristics and biases. Cambridge University Press, Cambridge

Kipper K, Dang HT, Stone Palmer M (2000) Class-Based Con- struction of a Verb Lexicon. In: Henry A. Kautz and Bruce W. Porter (eds.), Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, July 30-August 3, 2000, Austin, Texas, USA., AAAI Press/The MIT Press, 691–696

Kliegr T, Bahník Š, Fürnkranz J (2021) A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. Artif Intell 295:103458

Khodak M, Risteski A, Fellbaum C, Arora S (2017) Automated WordNet construction using word embeddings. In: J. Camacho-Collados and M. T. Pilehvar (eds.), Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications. Association for Computational Linguistics, 12–23, DOI: https://doi.org/10.18653/v1/W17-1902.

Konishi T (1993) The semantics of grammatical gender: a cross-cultural study. J Psycholinguist Res 22(5):519–534

Lai CK, Skinner AL, Cooley E, Murrar S, Brauer M, Devos T et al (2016) Reducing implicit racial preferences: II. Intervention effectiveness across time. J Expl Psychol Gen 145(8):1001

Leinbach MD, Fagot BI (1993) Categorical habituation to male and female faces: gender schematic processing in infancy. Infant Behav Dev 16(3):317–332

Lewis M, Lupyan G (2020) Gender stereotypes are reflected in the distributional structure of 25 languages. Nat Hum Behav 4:1021–1028

Mazzuca C, Santarelli M (2022) Making it abstract, making it contestable: politicization at the intersection of political and cognitive science. Rev Philos Phychol. PsyArXiv Preprint. https://doi.org/10.31234/osf.io/u6wd2

Mazzuca C, Majid A, Lugli L, Nicoletti R, Borghi AM (2020a) Gender is a multifaceted concept: evidence that specific life experiences differentially shape the concept of gender. Lang Cogn. https://doi.org/10.1017/langcog.2020.15

Mazzuca C, Borghi AM, van Putten S, Lugli L, Nicoletti R, Majid A (2020b) Gender at the interface of culture and language: conceptual variation between Italian, Dutch, and English. PsyArXiv Preprint. https://doi.org/10.31234/osf.io/dpa8s

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM Comput Surv (CSUR) 54(6):1–35

Minsky ML (1991) Logical versus analogical or symbolic versus connectionist or neat versus scruffy. AI Mag 12(2):34

Mikolov T, Yih WT, Zweig G (2013) Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies, 746–751

Mikolov T, Grave E, Bojanowski P, Puhrsch C, Joulin A (2017) Advances in pretraining distributed word representations. arXiv preprint arXiv:1712.09405

Misersky J, Gygax PM, Canal P, Gabriel U, Garnham A, Braun F et al (2014) Norms on the gender perception of role nouns in Czech, English, French, German, Italian, Norwegian, and Slovak. Behav Res Methods 46(3):8

Misersky J, Majid A, Snijders TM (2019) Grammatical gender in German influences how role-nouns are interpreted: evidence from ERPs. Discourse Process 56(8):643–654

Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. Big Data Soc 3(2):2053951716679679

Moss-Racusin CA, Dovidio JF, Brescoll VL, Graham MJ, Handelsman J (2012) Science faculty's subtle gender biases favor male students. Proc Natl Acad Sci 109(41):16474–16479

Mukerjee A, Biswas R, Deb K, Mathur AP (2002) Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management. Int Trans Oper Res 9(5):583–597

Narayanan S, Fillmore CJ, Baker CF Petruck MR (2000) Framenet meets the semantic web: a daml+ oil frame representation. Technology 2003

Navigli R, Ponzetto SP (2012) BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif Intell 193:217–250

Nielsen FÅ (2017) Wembedder: Wikidata entity embedding web service. arXiv preprint arXiv:1710.04099

Nissim M, van Noord R, van der Goot R (2020) Fair is better than sensational: man is to doctor as woman is to doctor. Comput Linguist 46(2):487–497

Nosek BA, Banaji MR, Greenwald AG (2002) Harvesting implicit group attitudes and beliefs from a demonstration web site. Group Dyn Theory Res Pract 6(1):101

Nuzzolese AG, Gangemi A, Presutti V (2011) Gathering lexical linked data and knowledge patterns from FrameNet. In: M.A. Musen and O. Corcho (eds), Proceedings of the 6th International Conference on Knowledge Capture (K-CAP 2011), June 26–29, 2011, Banff, Alberta, Canada, ACM, 2011, 41–48

Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453

Peng A, Nushi B, Kıcıman E, Inkpen K, Suri S, Kamar E (2019) What you see is what you get? The impact of representation criteria on human bias in hiring. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 7(1), 125–134

Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532–1543

Pereira F, Lou B, Pritchett B, Ritter S, Gershman SJ, Kanwisher N, Botvinick M, Fedorenko E (2018) Toward a universal decoder of linguistic meaning from brain activation. Nat Commun 9(1). https://doi.org/10.1038/s41467-018-03068-4

Poulin-Dubois D, Serbin LA, Kenyon B, Derbyshire A (1994) Infants' intermodal knowledge about gender. Dev Psychol 30(3):436

Rebele T, Suchanek F, Hoffart J, Biega J, Kuzey E, Weikum G (2016) YAGO: a multilingual knowledge base from wikipedia, wordnet, and geonames. International semantic web conference. Springer, Cham, pp 177–185

Richeson JA, Shelton JN (2007) Negotiating interracial interactions: costs, consequences, and possibilities. Curr Dir Psychol Sci 16(6):316–320

Rippon G (2019) The Gendered Brain: the new neuroscience that shatters the myth of the female brain. The Bodley Head Ltd, London

Risman BJ (2004) Gender as a social structure: theory wrestling with activism. Gend Soc 18(4):429–450

Rothe S, Schütze H (2017) Autoextend: combining word embeddings with semantic resources. Comput Linguist 43(3):593–617

Rudman LA, Greenwald AG, McGhee DE (2001) Implicit self-concept and evaluative implicit gender stereotypes: self and ingroup share desirable traits. Pers Soc Psychol Bull 27(9):1164–1178

Saguy T, Reifen-Tagar M, Joel D (2021) The gender-binary cycle: the perpetual relations between a biological-essentialist view of gender, gender ideology, and gender-labelling and sorting. Philos Trans R Soc B 376(1822):20200141

Samuel S, Cole G, Eacott MJ (2019) Grammatical gender and linguistic relativity: a systematic review. Psychon Bull Rev 26(6):1767–1786

Sax L (2005) Why gender matters: what parents and teachers need to know about the emerging science of sex differences. Doubleday, New York

Shiffrin RM, Schneider W (1977) Controlled and automatic human information processing: II. Perceptual learning automatic attending and a general theory. Psychol Rev 84(2):127–190. https://doi.org/10.1037/0033-295X.84.2.127

Simon HA (1955) A behavioral model of rational choice. Quart J Econ 69:99–118

Speer R, Chin J, Havasi C (2017) ConceptNet 5.5: an open multilingual graph of general knowledge, In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI Press, 4444–4451

Strauss E, Sherman EMS, Spreen O, Spreen OA (2006) Compendium of neuropsychological tests: administration, norms, and commentary, 3rd edn. Oxford University Press, New York

Swinger N, De-Arteaga M, Heffernan IV NT, Leiserson MD, Kalai AT (2019) What are the biases in my word embedding?. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 305–311

Taylor MG, Rhodes M, Gelman SA (2009) Boys will be boys; cows will be cows: children's essentialist reasoning about gender categories and animal species. Child Dev 80:461–481

Treviño LJ, Gomez-Mejia LR, Balkin DB, Mixon FG Jr (2018) Meritocracies or masculinities? The differential allocation of named professorships by gender in the academy. J Manag 44(3):972–1000

Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases. Science 185(4157):1124–1131

van Anders SM (2015) Beyond sexual orientation: integrating gender/sex and diverse sexualities via sexual configurations theory. Arch Sex Behav 44(5):1177–1213

van Assem M, Gangemi A, Schreiber G (2006) Conversion of Wordnet to a standard RDF/OWL representation. In: Proceedings of the fifth international conference on language resources and evaluation (LREC'06)

van Loon A, Freese J (2019) Word embeddings reveal how fundamental sentiments structure natural language. SocArXiv Preprint. https://doi.org/10.31235/osf.io/r7ewx

van Putten S, O'Meara C, Wartmann F, Yager J, Villette J, Mazzuca C, Bieling C, Burenhult N, Purves R, Majid A (2020) Conceptualisations of landscape differ across European languages. PLoS One 15(10):e0239858

West C, Zimmerman DH (1987) Doing gender. Gend Soc 1(2):125–151

West M, Kraut R, Ei Chew H (2019) I'd blush if I could: closing gender divides in digital skills through education. Technical report. Unesco and Equals, https://unesdoc.unesco.org/ark:/48223/pf0000367416

Zhai M, Tan J, Choi JD (2016) Intrinsic and extrinsic evaluations of word embeddings. In: D. Schuurmans and M. Wellman (eds), Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence AAAI Press, 4282–4283

Zhao J, Wang T, Yatskar M, Cotterell R, Ordonez V, Chang KW (2019) Gender bias in contextualized word embeddings. arXiv preprint arXiv:1904.03310

Zhang Y, Han K, Worth R, Liu Z (2020) Connecting concepts in the brain by mapping cortical representations of semantic relations. Nat Commun 11(1). https://doi.org/10.1038/s41467-020-15804-w

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.