Social biases in word embeddings and their relation to human cognition

Aylin Caliskan[1] and Molly Lewis[2,3]

[1]Department of Computer Science, George Washington University

[2]Department of Psychology, Carnegie Mellon University

[3]Department of Social and Decision Sciences, Carnegie Mellon University

Author Note

Authors listed in alphabetical order.

Social biases in word embeddings and their relation to human cognition

Introduction

Infants come into the world ready to learn about their social environment, and they do so quickly (Tasimi, 2020). By the time they reach school, children show evidence of holding common gender and racial stereotypes. For example, in one classic demonstration of children's stereotypes, children are asked to "draw a picture of a scientist" (Chambers, 1983). About 60% of kindergartners will draw a male scientist given these instructions and, by the time they reach high school, about 80% of all students will (Miller et al., 2018). Behavioral tasks such as these demonstrate that children develop increasingly adult-like stereotypes of the social world over time, and there is reason to think that these stereotypes may ultimately contribute to structural inequality (Charlesworth & Banaji, 2019a; Huang et al., 2020). An obvious question to ask, then, is where do these stereotypes come from?

Language is one likely source. Language is a particularly powerful means of communicating information because it can convey a message without direct experiences — a child can acquire the stereotype that most scientists are men by hearing statements about scientists, but without ever encountering a real, live scientist. Information about the world is conveyed through language in multiple ways, such as through explicit statements. For instance, a child who hears the statement, "Men make better scientists than women," might be inclined to develop a stereotype that men are more suited for life as a scientist, while women should pursue other careers. Experimental research suggests that children are able to learn about social stereotypes through various types of explicit linguistic statements (Bian et al., 2017; Cimpian & Markman, 2011; Cimpian et al., 2012; Rhodes et al., 2019).

The focus of this chapter is on a second, less studied route through which language conveys social meaning: the co-occurrence of words in a large corpus of language. This route is often referred to as "distributional statistics" or "distributional semantics" (in this chapter, we use the latter; Harris, 1951, 1954; Lenci, 2008). The intuition underlying

distributional semantics is that words that occur in similar linguistic contexts have similar meanings. Distributional semantics would tell you, for example, that “plate” and “bowl” have similar meanings because they both occur in similar linguistic contexts. The core idea of this approach—that meanings are defined by their co-occurrence statistics—has its roots in structuralist thought. The central idea of structuralism is that language can be thought of as an isolated system, and meaning in the system can be derived by considering only the relationships between units within the system (Saussure, 1916, 1960). Distributional semantics is a modern instantiation of this idea. Models of distributional semantics were first introduced by cognitive scientists in the 1990s (Landauer & Dumais, 1997; Lund & Burgess, 1996) and, since then, the machine learning community has developed these models to be more sophisticated and accurate (Bengio et al., 2003; Bojanowski et al., 2016; Mikolov, Chen, et al., 2013; Pennington et al., 2014). These more modern instantiations of distributional semantics are often referred to as “word embeddings.”

Word embeddings are low-dimensional numeric representations of words generated by artificial intelligence (AI) methods that capture word co-occurrence statistics. The assumption in these models is that words located in close proximity to one another in the vector space are semantically similar. The similarity between two word meanings, such as “plate" and “bowl", can be quantified by taking the cosine distance between the corresponding vectors in the model. Word embedding models also capture more analogical similarity relations. For example, by applying simple arithmetic operations to word vectors, these models capture the fact that ‘Paris’ is to ‘France’ as ‘Rome’ is to ‘Italy’ (Mikolov, Sutskever, et al., 2013). State-of-the-art word embeddings have lead to advances in natural language processing and understanding tasks such as machine translation and human-like text generation (Bengio et al., 2003; Mikolov, Chen, et al., 2013; Pennington et al., 2014).

Critically, recent research suggests that word embeddings not only encode information about mundane meanings, like “plate” and “bowl” or relations such as “Paris” and “France” but they also encode associations with social import, such as “woman” and

"nurse" or "man" and "doctor" (Bolukbasi et al., 2016; Caliskan et al., 2017; Lewis & Lupyan, 2020). These biases are apparent in statistical machine translation systems using word embeddings to translate sentences (Bolukbasi et al., 2016). For example, these algorithms tend to translate a sentence like, 'One$_{neutral}$ is a nurse' from a gender-neutral language to a gendered language like English as 'She is a nurse', implicitly reflecting the bias that nurses tend to be female. Demonstrations such as this show that word embeddings encode subtle information about social biases, and when used to solve AI tasks, these embeddings may serve to themselves perpetuate biases.

The fact that information about social stereotypes is available in the input leads to an intriguing psychological hypothesis: a learner's exposure to language statistics that reflect a social stereotype might lead them to acquire that stereotype. That is, in much the same way that a child might learn the stereotype that scientists tend to be male from an explicit statement like "Men make better scientists than women," a child might learn that same stereotype from tracking the distribution of the word "scientist" and words directly or indirectly related to gender in a large corpus of (biased) text. We refer to this as the "causal embedding hypothesis" (Caliskan et al., 2017; Greenwald, 2017; Lewis et al., 2020).

It is worth commenting on the theoretical status of this hypothesis. Word embeddings provide a numerical representation of the information that one could in principle extract from tracking the co-occurrences of words in a large corpus of text, independent of physical constraints. This is similar to ideal observer analyses used in different domains in psychology (Geisler, 2003), where the goal is to quantify how an optimal learner (in a Bayesian sense) would behave in a task if they had all available information and did not have cognitive limitations. The causal embedding hypothesis is not a Bayesian theory, but it is similar in the sense that the goal is to assess what information is available in the input, and whether the cognitive system uses that information. Critically, the causal embedding hypothesis is not a hypothesis about the mechanism that the cognitive system applies to this information; the cognitive system could use the same

information as the input to word embeddings to derive word meanings, applying a very different algorithm than that used by word embeddings. The causal embedding hypothesis is a claim only about whether humans make use of language co-occurrence statistics to derive word meanings, and is agnostic about the mechanism through which this is achieved (in Marr's (1982) terms, it is a "computational level theory").

There is reason to think, however, that children and adults have learning mechanisms that would allow them to track meanings from co-occurrence statistics. In particular, we know that infants are able to track statistics in their environment to learn information about language (see Saffran & Kirkham, 2018, for a review). For example, 8-month olds can learn to detect word boundaries from a stream of continuous speech by tracking the co-occurrences between sounds (Saffran et al., 1996). There is further evidence that the ability to track environmental statistics extends beyond language to the physical (e.g., Kirkham et al., 2002; Téglás et al., 2011) and social worlds (e.g., Johnson et al., 2007; Wellman et al., 2016). Taken together, this body of experimental work suggests that statistical learning is a powerful, general learning mechanism that is available to learners. It is plausible that children and adults could learn word meanings in part by tracking word co-occurrences in their input (see Günther et al., 2019, for further discussion).

Recently developed word embedding methods allow a key prediction of the causal embedding hypothesis to be tested—namely, that there should be a close correspondence between social biases in distributional semantics and those in social cognition. These methods reveal that word embeddings trained on large corpora of text reflect the same social biases that have been demonstrated behaviorally in the social psychology literature (Caliskan et al., 2017; Greenwald et al., 1998; Nosek et al., 2002a). Importantly, note that while these findings are a key data point for evaluating the causal embedding hypothesis, they are not determinate of it. Another hypothesis for explaining the close correspondence between language statistics and social biases is that there's a causal arrow in the other direction; that social biases shape language statistics. This alternative possibility is almost

certainly true, but it is not mutually exclusive with the causal embedding hypothesis. Surprisingly, despite its important implications, little work to date has directly tested the causal embedding hypothesis. We return to this point at the end of the chapter.

The goal of this chapter is to review evidence for the close correspondence between social biases in word embeddings and human cognition, and highlight fruitful areas of future research. We 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. We conclude by discussing ways for robustly generalizing the methods we describe to languages beyond English, and for testing the causal embedding hypothesis more directly.

Word embeddings reflect social biases in human cognition

To what extent do word embeddings reflect social biases in human cognition? We consider evidence that word embeddings encode information that goes beyond encyclopedic knowledge to meanings in the social domain, such as gender and race, in a way that accords with human judgments of semantic similarity. In order to assess the relationship between social biases in word embeddings and those in human cognition, we need a method for measuring human biases. We first describe the primary method we use for quantifying social biases in human cognition: the Implicit Association Test (IAT). We then describe a method for quantifying social biases in word embeddings in a way that closely aligns with the IAT, followed by several extensions of this method.

Quantifying social biases in human cognition

To quantify participants' social biases, we make use of a well-studied behavioral task developed by social psychologists, called the IAT (Greenwald et al., 1998). The IAT uses reaction time to measure participants' associations between two target concepts (e.g. flower vs. insect) and two target attributes (e.g. pleasant vs. unpleasant) in a word categorization task. Participants are presented with a single word corresponding to one of the concepts or

attributes (e.g., "rose" or "happiness"), and are asked to make a two-alternative categorization decision. In the compatible block of the test, the stereotypically associated concepts and attributes share a response key (e.g. flower/pleasant vs. insects/unpleasant); in the incompatible block, the non-stereotypically associated concepts and attributes share a response key (e.g. flower/unpleasant vs. insects/pleasant). Participants tend to categorize a word more quickly in the compatible block, relative to the incompatible block, and this pattern is taken as evidence for a closer cognitive association between the compatible concept-attribute meanings, relative to the incompatible concept-attribute meanings. For example, participants tend to respond more quickly when the response key for flower words is the same as that for pleasant words (vs. unpleasant words), suggesting that participants have a cognitive association between flowers and pleasantness.

A classical Cohen's d effect size can be calculated for group-level performance in the IAT to quantify bias. Let $M_{incompatible}$ and $M_{compatible}$ be the log-transformed, mean reaction time in the incompatible and compatible blocks, respectively. Let the pooled standard deviation be the mean of $SD_{incompatible}$ and $SD_{compatible}$. Cohen's d can then be calculated as:

$$d = \frac{M_{incompatible} - M_{compatible}}{\text{Pooled SD}} \tag{1}$$

A related IAT effect size measure, D score, computes the standard deviation in the denominator ignoring condition assignment (the standard deviation across all conditions; Greenwald et al., 2003).

A large body of literature has implemented the IAT with a range of stereotyped concept and attribute pairs in order to measure the strength of social biases (e.g., Kiefer & Sekaquaptewa, 2007; Nosek et al., 2002a; Stanley et al., 2011). For example, the IAT can be used to measure the extent to which participants hold the stereotype that women are more closely associated with family life, while men are more closely associated with career life. Estimates for the strength of various biases can be obtained from published studies on the IAT, typically from a sample of around 50 participants. In addition, a virtual

laboratory, called Project Implicit (https://implicit.harvard.edu/implicit/; Nosek et al., 2002a), has administered many different types of IATs to millions of people all over the world. These data are freely available to researchers to evaluate group level biases from diverse populations, with very large sample sizes. Importantly, Project Implicit allows the researcher to not just ask questions about groups of people from convenient samples (e.g., undergraduates at prestigious universities), but to compare different social groups (e.g., US participants vs. Mexican participants; participants with a college degree vs. those without). Data from both published studies and Project Implicit provide valuable estimates of bias in human cognition that can then be compared to analogous estimates in language statistics.

Quantifying social biases in language statistics

In order to compare the magnitude of biases in social cognition to those in language statistics, we need a principled method for quantifying social biases in word embeddings. Caliskan et al. (2017) introduced the Word Embedding Association Test (WEAT) which derives an estimate of bias in word embeddings that can be compared to the effect size of group-level human performance in the IAT. WEAT borrows the word stimuli representing the target social groups and evaluative attributes from the IATs designed by social psychologists, and then uses the distance between a pair of vectors (more precisely, their cosine similarity score) as analogous to reaction time in the IAT. The assumption is that semantic 'nearness' in the embedding space implies less reaction time in the IAT's pairing tasks (see Figure 1; McDonald & Lowe, 1998; Moss et al., 1995).

WEAT computes the differential association scores for each word in the target concept and attribute categories. The unit of association in WEAT, $s(w, A, B)$ (Equation 2), quantifies the association of an attribute word, w (e.g., "home") with each word from

the target categories, A (e.g. "male") and B (e.g. "female").

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b}) \tag{2}$$

Figure **??** summarizes the computation of a unit of association in WEAT. In Figure **??**, the target concepts are male and female, and the word "home" is an example stimulus word (from the "family" attribute). In this case, $s(w, A, B)$ measures the degree to which the word "home" is similar to male words relative to female words.

In order to derive an effect size measure comparable to the behavioral data, $s(w, A, B)$ is calculated for all attribute words for each of the two attribute categories. The effect size quantifies the differential association between two sets of target categories X, Y and two sets of evaluative attributes A, B learned by word embeddings, a standardized bias score analogous to Cohen's d effect size estimates in the IAT (Cohen, 2013). The mean association score for each word is then divided by the pooled standard deviation of association scores to obtain the WEAT effect size of bias $ES(X, Y, A, B)$:

$$ES(X, Y, A, B) = \frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std-dev}_{w \in X \cup Y} s(w, A, B)} \tag{3}$$

It is worth noting that these effect sizes don't have the same interpretation as the IAT, as the "participants" in WEAT experiments are words, not people. As a result, WEAT reflects the overall linguistic biases of individuals whose writing samples were included in the corpus that was used in training the word embeddings.

Caliskan et al. (2017) applied the WEAT to models trained on large corpora of English text. Strikingly, across eight different IAT types, ranging from attitudes about race to attitudes about people with disabilities, all resulted in positive effect sizes. This suggests the word embedding models encode many of the same social biases that humans do. Table 1 shows all the IAT types and corresponding WEAT findings for these stereotypes. Further, these findings generalize to word embeddings trained with various machine

learning algorithms on different corpora. For instance, applying WEAT to Word2vec's word embeddings trained on the Google News corpus generates similar results while reflecting the artefacts of the training set and algorithm. Similar IAT biases have also been found using less sophisticated distributional semantics algorithms (Bhatia, 2017).

Together, these findings suggest that word embeddings encode not just innocuous semantic relationships, like the relationship between "plate" and "bowl", they also encode social biases in a way that closely corresponds to biases in human cognition, as measured by the IAT. This pattern is consistent with the idea that people learn some information about social stereotypes from biases in language statistics.

Quantifying social biases in language statistics cross-linguistically

The IAT provides estimates of the degree to which people hold a particular psychological bias, and people vary substantially in the degree of bias they hold. This variability bias is not entirely random — it is often predicted by demographic factors of participants. For example, White American children (Newheiser & Olson, 2012) and adults (Kurdi et al., 2019; Nosek et al., 2002a) show a bias to associate Whites with positive meanings (e.g., "awful", "spider") and Blacks with positive meanings ("happy", "flowers"), whereas Black Americans show no such bias (or even the opposite pattern). While the sources of this difference are likely highly complex, the close relationship between word embeddings and implicit measures of biases suggests one possible causal source: The co-occurrence statistics of the language input to which a person is exposed shapes the strength and type of biases they hold. If Black and White Americans are in partially distinct communities, their language input will be different. More generally, if language statistics shape psychological bias, then communities with language that have more bias in their statistics, will also have speakers with more bias psychologically, as measured by tasks like the IAT.

Recent work by Lewis and Lupyan (2020) tests one version of this hypothesis by

Target words	**Attribute words**	**IAT Finding**				**WEAT Finding**			
		Ref	**N**	d	p	$\mathbf{N_T}$	$\mathbf{N_A}$	d	p
Flowers vs insects	Pleasant vs unpleasant	(Greenwald et al., 1998)	32	1.35	10^{-8}	25×2	25×2	1.50	10^{-7}
Instruments vs weapons	Pleasant vs unpleasant	(Greenwald et al., 1998)	32	1.66	10^{-10}	25×2	25×2	1.53	10^{-7}
Eur.-American vs Afr.-American names	Pleasant vs unpleasant	(Greenwald et al., 1998)	26	1.17	10^{-5}	32×2	25×2	1.41	10^{-8}
Male vs female names	Career vs family	(Nosek et al., 2002a)	$39k$	0.72	$< 10^{-2}$	8×2	8×2	1.81	10^{-3}
Math vs arts	Male vs female terms	(Nosek et al., 2002a)	$28k$	0.82	$< 10^{-2}$	8×2	8×2	1.06	.018
Science vs arts	Male vs female terms	(Nosek et al., 2002b)	91	1.47	10^{-24}	8×2	8×2	1.24	10^{-2}
Mental vs physical disease	Temporary vs permanent	(Monteith & Pettit, 2011)	135	1.01	10^{-3}	6×2	7×2	1.38	10^{-2}
Young vs old people's names	Pleasant vs unpleasant	(Nosek et al., 2002a)	$43k$	1.42	$< 10^{-2}$	8×2	8×2	1.21	10^{-2}

Table 1

Summary of Word Embedding Association Tests (WEAT) replicating 8 well-known IAT findings using word embeddings, using GloVe word embeddings (Caliskan et al., 2017). Each result compares two sets of words from target concepts *about which we are attempting to learn with two sets of* attribute *words. In each case, the first target is found compatible with the first attribute, and the second target with the second attribute, following a stereotype congruent order. Throughout, the word lists are collected from the studies that are being replicated. N: number of subjects. N_T: number of target words. N_A: number of attribute words. The results are reported in effect sizes (d) and p-values (p, rounded up) to emphasize that the statistical and substantive significance of both sets of results is uniformly high; however these results do not necessarily imply that the results are directly comparable to those of human studies. For the online IATs (rows 4, 5, and 8), p-values were not reported, but are known to be below the significance threshold of* 10^{-2}.

examining the language statistics of distinct language communities in the most extreme form: speakers of different languages. Lewis and Lupyan applied the career-gender WEAT to word embedding models trained on 25 languages. Several modifications were made to the WEAT in order to apply it to languages beyond English. First, the word stimuli were translated from English into the sample of 25 languages by native speakers. In the original career-gender IAT, the gender words were proper names ("John", "Amy"), but because there are no direct translation equivalents of proper names, alternate words were used that directly denoted gender (e.g., "man", "woman"). Second, for languages with grammatical gender, some of the target words had multiple forms. For example, in Spanish, "niños" refers to male children and "niñas" refers to female children. To address this issue, word distances were calculated for same gender pairs only (e.g., children-male ("niños") to man-male ("hombre"), and children-female ("niños") to man-female ("mujer")).

The authors found that almost all of the languages (22 of 25) had a bias to associate women with home and males career, but there was a large amount of variability across languages in the degree of this bias. Does this variability predict cross-linguistic differences in psychological gender bias?

To answer this question, the degree of psychological career-gender bias was estimated for speakers of each of the target languages. Estimates were obtained from a large dataset of over 650 thousand gender-career IATs administered world wide to speakers of each of the 25 languages (Nosek et al., 2002a). For each language, a mean IAT score was calculated across all speakers of the target language. Like for the language measure, there was substantial variability across speakers of different languages in the degree to which they held a psychological bias to associate men with the concept of career and women with the concept of home. For instance, participants from Mexico and Sweden had a relatively weak bias, whereas the participants from the Netherlands and Brazil had a relatively strong bias. Critically, the strength of the bias in the statistics of a speaker's language was moderately correlated with the strength of the psychological bias for speakers of that language (Fig. 3).

The cross-linguistic career-gender WEAT has also been shown to be related to other measures of psychological bias, beyond the career-gender IAT. In particular, the magnitude of the career-gender bias in a language is related to two psychological constructs that have been previously claimed to be explanations of structural gender inequality. The first is a measure that quantifies the extent to which genders differ in their general preferences (e.g., willingness to take risks, patience, etc.; Falk & Hermle, 2018). The gender-career WEAT is moderately correlated with this measure, such that countries with greater differences in gender preferences also have greater gender bias present in their languages. Second, previous research has argued that there are gender differences in the degree to which members of different genders have a sense of "self-efficacy" in science, technology, engineering, and mathematics (STEM) fields (Stoet & Geary, 2018). Lewis and Lupyan (2020) find that countries with greater gender differences in self-efficacy also have greater gender bias present in their languages. These findings demonstrate that the correspondence between gender bias in language statistics and psychological gender bias generalizes beyond one particular measure of psychological gender bias, and suggests that language statistics may have broad explanatory power in accounting for psychological constructs.

In sum, these data points are consistent with the causal embedding hypothesis, but are not conclusive since the design is correlational and thus unable to establish causality. Nevertheless, establishing that the correspondence between human judgments and language statistics generalizes beyond English is an important first step to establishing causality.

Quantifying social biases in linguistic input to children

If the causal embedding hypothesis is correct, then we should expect social biases to be present in the linguistic input to people who are beginning to learn the social biases in their culture: children.

Several recent studies have evaluated the extent to which social biases are present in the statistics of linguistic input to children. Lewis et al. (2020) trained word embedding

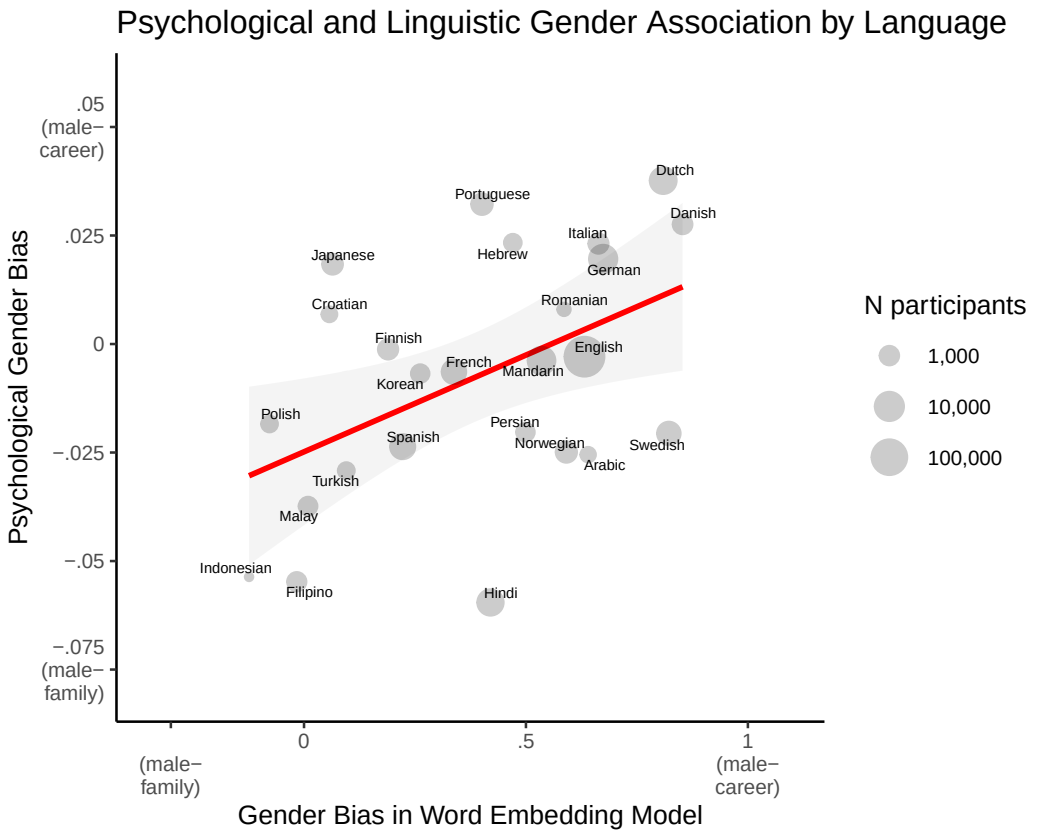

Figure 3

Data from Lewis and Lupyan (2020) showing the cross-linguistic relationship between the magnitude of the career-gender bias in the word embedding model trained on native language text, and native speakers' degree of psychological bias as measured by the Implicit Association Test (IAT). These results suggest that speakers of languages with greater bias in the language statistics tend to have greater psychological bias.

models on the text from a corpus of popular, contemporary books read to young children (0-5 years), and measured the magnitude of several biases in the corpus using the WEAT. They found evidence for three biases present in the statistics of the children's books corpus: a bias to associate males with career and females with family, a bias to association males with math and females with language, and a bias to associate males with math and females with art. Each of these biases have been demonstrated behaviorally using the IAT, some of them with children. Strikingly, the magnitudes of the biases in the language statistics were larger in the children's book corpus, relative to models trained on comparably sized adult fiction. This pattern suggests that children's books may contain

exaggerated representations of stereotypes to children, making them a potentially powerful means through which stereotypes are transmitted.

Charlesworth et al. (2020) examined a similar set of gender biases in language using the WEAT, across a diverse set of corpora. They analyzed embedding models trained on naturalistic child-directed speech, child-produced speech, historical children's books, and transcripts from child-directed movies and TV shows. Their data show consistent biases across corpora (meta-analytic estimate of bias: $D = 0.57$), and show roughly equal degrees of bias in child-directed and adult-directed corpora (in contrast to Lewis et al., 2020). They also go beyond stereotypes that have been well-documented in the behavioral literature, and ask whether a wide range of other stereotypes are also present the language statistics of children's input. They find, for example, that across corpora, females are associated with the traits "shy", "affectionate" and "gentle" whereas males are associated with the traits "direct", "tough" and "defensive."

This set of findings provides evidence that gender biases are present in the language statistics of input to children, possibly to a greater degree than in adult-directed input, suggesting that language statistics could provide children with information that could shape their early social biases.

Word embeddings reflect structural measures of bias

So far, we have presented evidence that there is a correspondence between social biases in language statistics and those in human cognition, as measured by the WEAT and the IAT. Next, we ask whether there is also evidence that language statistics are related to objective, structural measures of bias. For example, are occupations that are more female biased in language statistics also more female biased in the actual workforce? There is already independent evidence that group-level measures of psychological bias are associated with structural inequality. For example, Nosek et al. (2009) find that countries with greater gender-science stereotype have greater gender inequality in science achievement. Evidence

for a relationship between language statistics and objective measures of bias could suggest a causal pathway whereby language statistics shape psychological biases, and psychological biases in turn lead to structural inequality.

The Word Embedding Factual Association Test (WEFAT; Caliskan et al, 2017) provides one method for comparing language statistics to objective measures of bias. The WEFAT is a standardized effect size measure of the difference in two distributions of associations between a single target word and two sets of attributes. Equation 4 below presents the WEFAT formally, using the same notation conventions as Equation 3.

$$ES(\vec{w}, A, B) = \frac{\text{mean}_{a \in A}\cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B}\cos(\vec{w}, \vec{b})}{\text{std-dev}_{x \in A \cup B}\cos(\vec{w}, \vec{x})} \tag{4}$$

The WEFAT differs from the WEAT in that it measures the association of a single target word to the evaluative attribute sets, rather than comparing two sets of target words and two sets of attribute words (analogous to the single category IAT, SC-IAT, Karpinski and Steinman (2006)). It can be used to measure, for example, the extent to which the actual real world gender bias of an occupation (e.g., nurse) corresponds to biases in language statistics.

Caliskan et al. (2017) used the WEFAT to calculate a gender bias score for a set of 50 frequent occupation words (e.g., nurse, scientist, mechanic). They then compared the WEFAT score for each occupation to the actual labor force gender distribution using data from U.S. Bureau of Labor Statistics, by estimating the proportion of people in each occupation that were women. These two measures were strongly correlated with each other ($\rho = 0.90$; Figure 4), suggesting a close correspondence between bias in language statistics and structural inequality.

Using a measure similar to the WEFAT, Garg et al. (2018) replicate the relationship between objective occupation bias and language occupation bias demonstrated by Caliskan et al. (2017). Critically, they also find evidence for a close historical correspondence across

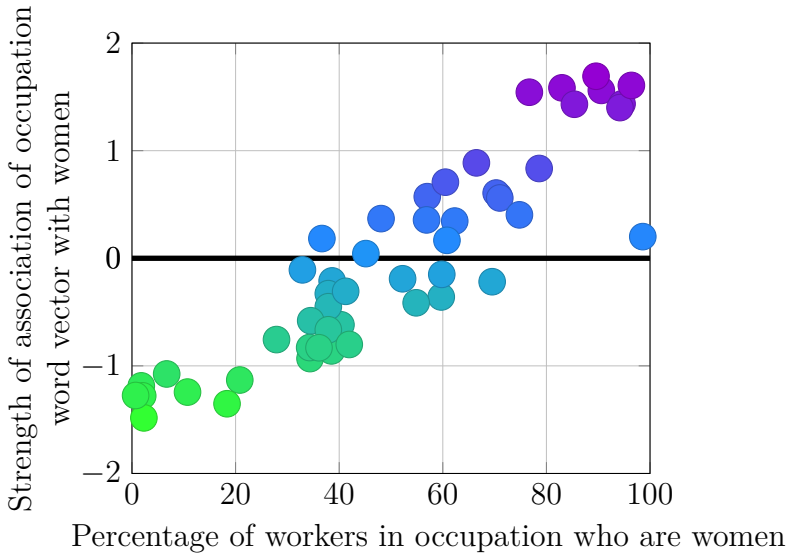

Figure 4

Occupation-gender association data from Caliskan et al. (2017). Pearson's correlation coefficient $\rho = 0.90$ (p-value $< 10^{-18}$) of proportion of women in 50 popular occupations retrieved from the annual reports of the U.S. Bureau of Labor Statistics and WEFAT gender scores for the corresponding 50 occupation names.

a 100 year period between the degree of gender bias for occupations in embeddings and in objective gender bias, as measured by census data. Their data show a historical trend for occupations to become, on average, less gender biased over time in both the embeddings and objective data (Figure 5). They find a similar pattern for race bias. Importantly, Garg et al. (2018) also directly examine the relationship between bias in word embeddings, objective bias, and psychological bias. Their data suggest that bias in word embeddings captures psychological bias, over and above what can be explained by objective bias.

The relationship between objective bias measures and word embeddings has also been examined at the language level. Lewis and Lupyan (2020) compared the degree of gender bias in the language spoken in a country, based on the career-gender WEAT, and the percentage of women among STEM graduates in tertiary education in that country. They found a correlation between the two: Countries that tended to have more women in STEM fields tended to have less bias in their language statistics. Further, there was also a

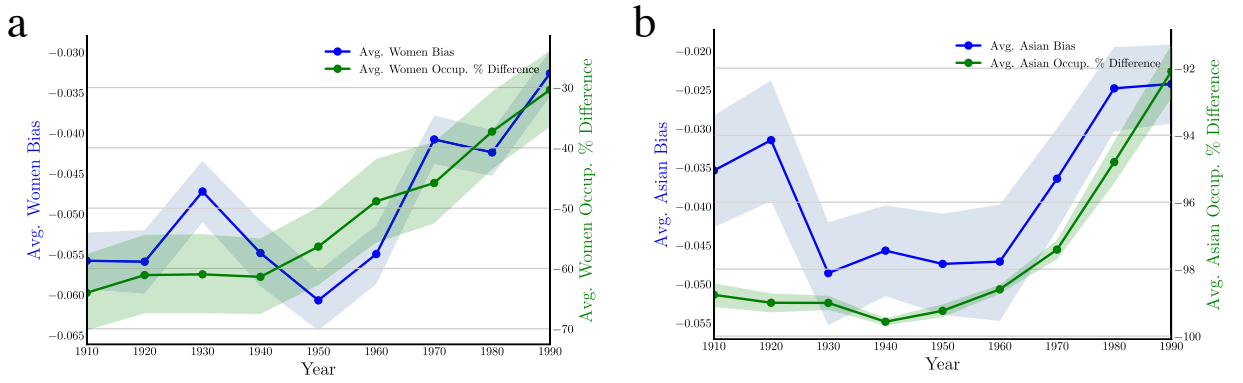

Figure 5

Data from Garg et al. (2018) showing the close correspondence in the strength of demographic biases in occupations between word embeddings and actual participation in the workforce over historical time. Panel a shows female bias and Panel b shows Asian bias. Y-axis shows the average bias across 76 different occupations (Blue: average embedding bias; Green: average bias toward women/Asian participation in occupations). Error bands show bootstrapped standard errors. Figures reproduced with permission from Garg, Schiebinger, Jurafsky, and Zou, 2018 (PNAS).

.

relationship between the degree of psychological gender bias in a country, as measured by the career-gender IAT, and female participation in STEM fields.

Together, these data raise the possibility that bias in language statistics could play a role in contributing to structural inequality. For example, consider the phenomenon of stereotype threat, whereby a person's knowledge that a negative stereotype applies to them leads to under-performance, particularly in underrepresented and intersectional group members in STEM (Spencer et al., 1999; Steele & Aronson, 1995). Bias in language statistics could be contributing to individuals' knowledge of these stereotypes, thereby exaggerating the effects of stereotype threat. Moreover, biased AI systems, such as the ones using word embeddings, are making consequential decisions about humans, such as college admissions and job candidate selection. As a result, gender bias in society that is learned by AI systems might become an additional factor contributing to gender inequity in the STEM workforce. Isolating the individual components of the bias lifecycle, that increasingly involves AI systems, can shed light on the causal factors contributing to mitigating, perpetuating, and amplifying bias in society as well as AI.

Future directions

Finally, we consider two important areas of future research for understanding the relationship between social biases in word embeddings and human cognition: Robust cross-linguistic generalization and approaches for testing the causal embedding hypothesis more directly.

Cross-linguistic generalization

An important area for future research is generalizing the approaches described in this chapter to other languages in a way that is maximally robust to cross-linguistic differences in morphosyntactic structure. Lewis and Lupyan (2020) were able to demonstrate cross-linguistic differences in bias using a course method for dealing with variability in the grammatical encoding of gender (by averaging vectors), but more robust methods are likely to more accurately capture biases in a diverse set of languages. In a related study, DeFranza et al. (2020) examined whether languages that encode grammatical gender have greater bias in their language statistics, rather than in psychological biases of speakers of those languages. They found that languages that encode gender grammatically tend to have a stronger positive association with men, relative to women, compared to languages that did not encode gender grammatically. However, it is not clear how much of these findings are due to grammatical vs. semantic associations without isolating the grammatical gender vector in the embedding space. Toney and Caliskan (2020) applied WEAT to seven languages, from five branches of varying language families, and show that word embeddings capture grammatical gender along with gender bias. Consequently, generating an accurate measure of linguistic bias in grammatically gendered languages requires isolating the gender vector. For example, when applying the gender-science WEAT in Polish by using the IAT words on Poland's Project Implicit, the resulting effect size signals stereotype-incongruent associations. Further analysis of this anomaly revealed that most of the words representing science in the Polish IAT have nouns

with feminine grammatical gender. However, when the gender direction is precisely identified and removed from the word embeddings while performing WEAT, the results are in line with the stereotype-congruent biases reported via IATs on the Project Implicit site (Nosek et al., 2009). These findings suggest that structural properties of languages should be taken into account when performing bias measurements on word embeddings.

Among social biases, so far, gender bias is the only one recognized as being related to a language's structural properties. Further analysis of languages might uncover more grammatical associations that might be captured while measuring certain types of biases. Moreover, the structure of a language might be causing limitations that don't allow for linguistic regularities to capture various associations. For example, Turkish does not have grammatical gender, and all the pronouns are gender-neutral. When measuring gender-science bias in Turkish word embeddings, the stereotype-incongruent results are not in agreement with IAT scores reported on Project Implicit (Nosek et al., 2002a). These unexpected results require in depth analysis. One potential reason might be the quality of word embeddings in Turkish. The low quality of the word embeddings might be due to the language-dependent pre-processing strategies that were not applied before training the embeddings. If the same types of pre-processing methods are applied to all languages from different language families, the training corpora might be losing important linguistic information. Another reason could be the fact that a training corpus in Turkish, which is a gender-neutral language, does not have a gender signal significant enough to capture gender associations. In Turkish, explicitly mentioning the gender of a subject requires including words such as 'kadın', 'erkek', 'kız', 'oğlan' which translate to English as 'woman/female', 'man/male', 'girl', 'boy'. Unless someone wants to emphasize the gender of a subject, for example when giving a stereotype-incongruent example such as 'O bir kadın doktor.' ('She is a doctor' in English), gender is not specified and it is instead inferred.

Generalizing WEAT to other languages requires taking the structure of a language into account. Understanding the relationship between linguistic structure and linguistic

bias can also help understand the causal factors behind learning biases in society.

Testing the causal embedding hypothesis directly

We have presented in this chapter a substantial body of evidence that psychological and structural measures of bias closely correspond to biases found in word embeddings. Where does this leave the causal embedding hypothesis?

The extant evidence for the causal embedding hypothesis is almost exclusively correlational and, as such, cannot provide strong evidence that biases in language statistics *shape* biases in human cognition and ultimately contribute to structural biases. There are consequently a range of possible causal models that are consistent with the data (e.g., human cognition shapes language statistics, but not the other way around, or a third variable shapes both human cognition and language statistics). Nevertheless, the goal is to build a causal theory (Grosz et al., 2020). Moving forward, there are two promising avenues for examining the extent to which language statistics play a causal role in shaping biases in human cognition.

The first is building causal models from observational data. Testing the causal embedding hypothesis using observational data is challenging because of the inextricable relationship between human cognition and language statistics: language statistics come from human minds. There are a range of statistical methods for inferring causality from observational data that are widely used in fields accustomed to inferring causality from observational data (e.g., political science, sociology, and economics; Grosz et al., 2020), and the rise of large-scale data in psychology, like Project Implicit, makes these methods more feasible (Lupyan & Goldstone, 2019; Paxton & Griffiths, 2017). One promising observational approach is to examine the relationship between language statistics and biases in human cognition over historical time. Evidence that biases in language statistics *precede* parallel change in biases in human cognition would provide some evidence for the causal embedding hypothesis. There is already compelling data to suggest there is

substantial change in biases in language statistics (Garg et al., 2018) and human cognition (Charlesworth & Banaji, 2019b) over historical time, but no work to date has examined the longitudinal relationship between the two.

The second approach for more directly testing the causal embedding hypothesis is experimentation. Experiments allow the researcher to randomly assign participants to conditions, and directly intervene on the causal system. Random assignment is the gold standard for demonstrating causal influence because it eliminates the possibility that background factors (i.e., confounds) are responsible for the observed effect. The experimental method can be used to test the causal embedding hypothesis by manipulating the statistics of participants' linguistic input and measuring their resulting psychological biases. Little work to date has taken this approach (but see McDonald & Ramscar, 2001), in part because the amount of linguistic input needed to change people's cognitive associations may be more than is practical to expose people to over the course of a short lab experiment (though this is an open empirical question). Future work could aim to manipulate the type of linguistic input people are exposed to over much longer timescales using experimental or quasi-experimental designs.

In applying these methods to understand the underlying causal dynamics, the resulting answer almost certainly will not be binary ("language statistics do shape human cognition"), but rather depend on many other aspects of human cognition. For example, does the strength of the effect of language statistics on human cognition change across development? Does the source of the language statistics matter? One possibility is that people are more influenced by language produced by a knowledgeable or respected speaker versus language produced by someone who is not perceived as knowledgeable or respected (Lewis & Frank, 2016; Xu & Tenenbaum, 2007). Another open question is whether some semantic domains are more influenced by language statistics than others. There is evidence now for a close correspondence between human cognition and language statistics for both encyclopedic knowledge (e.g., 'plate' and 'bowl') and social knowledge (e.g., 'women' are

more closely associated with 'home', than 'men'), but little is known about the differences in the relationship between language statistics and human cognition as a function of semantic domain. These questions, and others, wide open for future research.

Finally, there is an open question about the effect of language statistics at different scales. The fact that biases in language statistics correspond not only to biases in human cognition, but also to objective, structural measures of bias, suggests that language statistics may have effects that propagate beyond the individual mind. The causal dynamics between biases in language statistics, individual minds, biases in groups, and structural biases may vary (Anderson, 1972; Payne et al., 2017), and are a ripe area for future research. The wealth of available large-scale data makes answering questions about the causal dynamics at scales beyond the individual mind newly possible.

Conclusion

How do people learn stereotypes? In this chapter, we consider the possibility that biased associations in language statistics could be one contributing factor. We describe recently developed methods for measuring social biases in language statistics using word embedding models (WEAT and WEFAT), and review evidence that these measures closely correspond to both psychological measures of bias, as measured by the IAT, and more objective measures of bias, like the gender distribution of participation in the workforce. Notably, these biases are present not only in generic English corpora, like Wikipedia and Google News, but also in linguistic input to children and in multi-lingual corpora. The empirical evidence we present in this chapter is consistent with the causal embedding hypothesis, but there are many remaining open questions. Throughout, we highlight important research directions at the intersection of AI bias, bias in human cognition, computational linguistics, and language acquisition.

The methods we describe allow researchers to uncover, quantify, and validate implicit psychological biases and structural biases in an automated way, opening up a wide

space of future computer and information science as well as social science research. While we have focused on research by psychologists and computer scientists, researchers from many fields—linguists, sociologists, sociolinguists, neuroscientists, political scientists, philosophers, and policy-makers—can apply these methods to study the evolution of biases associated with populations across time and languages. For example, unlike the IAT that requires human participants, WEAT can be applied to historical corpora to study extinct languages or societies from centuries ago (Kozlowski et al., 2019; Toney & Caliskan, 2020). Further, while the focus of this chapter has been primarily on gender and racial biases, the methods we describe are generalizable to any imaginable stereotype.

In addition to the theoretical implications for social science, understanding how social biases found in language relate to human cognition has immediate, important practical applications. AI models likely perpetuate existing biases by generating biased outcomes, and so there is an urgent need to understand these causal dynamics so that language and AI can be used to mitigate potentially harmful biases.

References

Anderson, P. W. (1972). More is different. *Science*, *177*(4047), 393–396.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*(Feb), 1137–1155.

Bhatia, S. (2017). The semantic representation of prejudice and stereotypes. *Cognition*, *164*, 46–60.

Bian, L., Leslie, S.-J., & Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, *355*(6323), 389–391.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings, In *Advances in Neural Information Processing Systems*.

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

Chambers, D. W. (1983). Stereotypic images of the scientist: The draw-a-scientist test. *Science Education*, *67*(2), 255–265.

Charlesworth, T. E., & Banaji, M. R. (2019a). Gender in science, technology, engineering, and mathematics: Issues, causes, solutions. *Journal of Neuroscience*, *39*(37), 7228–7243.

Charlesworth, T. E., & Banaji, M. R. (2019b). Patterns of implicit and explicit attitudes: I. long-term change and stability from 2007 to 2016. *Psychological Science*, *30*(2), 174–192.

Charlesworth, T. E., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2020). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of 65+ million words.

Cimpian, A., & Markman, E. M. (2011). The generic/nongeneric distinction influences how children interpret new information about social others. *Child Development*, *82*(2), 471–492.

Cimpian, A., Mu, Y., & Erickson, L. C. (2012). Who is good at this game? Linking an activity to a social category undermines children's achievement. *Psychological Science*, *23*(5), 533–541.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

DeFranza, D., Mishra, H., & Mishra, A. (2020). How language shapes prejudice against women: An examination across 45 world languages. *Journal of Personality and Social Psychology*.

Falk, A., & Hermle, J. (2018). Relationship of gender differences in preferences to economic development and gender equality. *Science*, *362*(6412).

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644.

Geisler, W. S. (2003). Ideal observer analysis. *The Visual Neurosciences*, *10*(7), 12–12.

Greenwald, A. G. (2017). An ai stereotype catcher. *Science*, *356*(6334), 133–134.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197.

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives in Psychological Science*.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006–1033.

Harris, Z. S. (1951). Methods in structural linguistics.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162.

Huang, J., Gates, A. J., Sinatra, R., & Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*, *117*(9), 4609–4616.

Johnson, S. C., Dweck, C. S., & Chen, F. S. (2007). Evidence for infants' internal working models of attachment. *Psychological Science*, *18*(6), 501–502.

Karpinski, A., & Steinman, R. B. (2006). The single category implicit association test as a measure of implicit social cognition. *Journal of Personality and Social Psychology*, *91*(1), 16.

Kiefer, A. K., & Sekaquaptewa, D. (2007). Implicit stereotypes and women's math performance: How implicit gender-math stereotypes influence women's susceptibility to stereotype threat. *Journal of Experimental Social Psychology*, *43*(5), 825–832.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42.

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, *84*(5), 905–949.

Kurdi, B., Mann, T. C., Charlesworth, T. E., & Banaji, M. R. (2019). The relationship between implicit intergroup attitudes and beliefs. *Proceedings of the National Academy of Sciences*, *116*(13), 5862–5871.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, *20*(51).

Lewis, M., Cooper Borkenhagen, M., Converse, E., Lupyan, G., & Seidenberg, M. S. (2020). What might books be teaching young children about gender? https://psyarxiv.com/ntgfe

Lewis, M., & Frank, M. C. (2016). Understanding the effect of social context on learning: A replication of Xu and Tenenbaum (2007b). *Journal of Experimental Psychology: General*, *145*(9), e72–e80.

Lewis, M., & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behavior*, *4*, 1021–1028.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research methods, Instruments, & Computers*, *28*(2), 203–208.

Lupyan, G., & Goldstone, R. L. (2019). Introduction to special issue. beyond the lab: Using big data to discover principles of cognition. *Behavior Research Methods*, *51*(4), 1473–1476.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

McDonald, S., & Lowe, W. (1998). Modelling functional priming and the associative boost, In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. LEA.

McDonald, S., & Ramscar, M. (2001). Testing the distributioanl hypothesis: The influence of context on judgements of semantic similarity, In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality, In *Advances in neural information processing systems*.

Miller, D. I., Nolla, K. M., Eagly, A. H., & Uttal, D. H. (2018). The development of children's gender-science stereotypes: A meta-analysis of 5 decades of US draw-a-scientist studies. *Child Development*, *89*(6), 1943–1955.

Monteith, L. L., & Pettit, J. W. (2011). Implicit and explicit stigmatizing attitudes and stereotypes about depression. *Journal of Social and Clinical Psychology*, *30*(5), 484–505.

Moss, H. E., Ostrin, R. K., Tyler, L. K., & Marslen-Wilson, W. D. (1995). Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *21*, 863–883.

Newheiser, A.-K., & Olson, K. R. (2012). White and black american children's implicit intergroup bias. *Journal of Experimental Social Psychology*, *48*(1), 264–270.

Nosek, B. A. Et al. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, *106*(26), 10593–10597.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002a). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 101.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002b). Math= male, me= female, therefore math≠ me. *Journal of Personality and Social Psychology*, *83*(1), 44.

Paxton, A., & Griffiths, T. L. (2017). Finding the traces of behavioral and cognitive processes in big data and naturally occurring datasets. *Behavior Research Methods*, *49*(5), 1630–1638.

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*, *28*(4), 233–248.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation, In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*.

Rhodes, M., Leslie, S.-J., Yee, K. M., & Saunders, K. (2019). Subtle linguistic cues increase girls' engagement in science. *Psychological Science*.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, *69*.

Saussure, F. (1916, 1960). Course in General Linguistics. London, Peter Owen.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28.

Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences*, *108*(19), 7710–7715. https://doi.org/10.1073/pnas.1014345108

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797.

Stoet, G., & Geary, D. C. (2018). The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science*, *29*(4), 581–593.

Tasimi, A. (2020). Connecting the dots on the origins of social knowledge. *Perspectives on Psychological Science*, *15*(2), 397–410.

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, *332*(6033), 1054–1059.

Toney, A., & Caliskan, A. (2020). Valnorm: A new word embedding intrinsic evaluation method reveals valence biases are consistent across languages and over decades, arXiv 2006.03950. https://arxiv.org/abs/2006.03950

Wellman, H. M., Kushnir, T., Xu, F., & Brink, K. A. (2016). Infants use statistical sampling to understand the psychological world. *Infancy*, *21*(5), 668–676.

Xu, F., & Tenenbaum, J. B. (2007). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, *10*(3), 288–297.