Decomposing Implicit Bias in Distributional Semantic Models: The Roles of First- and Second-Order Co-Occurrence

Molly Apsel (mapsel@iu.edu)
Department of Psychological and Brain Sciences
Cognitive Science Program
Indiana University

Michael N. Jones (jonesmn@iu.edu)
Department of Psychological and Brain Sciences
Cognitive Science Program
Indiana University

Abstract

Distributional semantic models (DSMs) are computational models that learn semantic relationships through word co-occurrence patterns, broadly aligning with human statistical learning mechanisms. Prior research has shown that DSMs capture not only general semantic structure but also human social biases. For example, Caliskan et al. (2017) demonstrated that pre-trained word embeddings encode associations that mirror implicit stereotypes measured by the Implicit Association Test (IAT). To better understands how DSMs acquire these biases, we examined the roles of two distinct sources of distributional information: first-order (direct co-occurrence) and second-order (indirect co-occurrence) statistics. Our analysis revealed that nearly all biases tested could be accounted for by first-order statistics alone, while about half were significant in second-order statistics. Every bias was present in at least one of these co-occurrence types, with nuanced variation in how different topics exhibited bias across first- and second-order associations. These findings suggest that implicit biases in DSMs can be attributed to simple co-occurrence patterns, predominantly direct associations. Moreover, they support theories positing that implicit biases reflect statistical regularities in the environment rather than personal attitudes. This work highlights how these biases are embedded in natural language and how a cognitive system capable of statistical learning could acquire implicit biases through the same mechanisms that shape human semantic memory.

Keywords: distributional semantics; implicit bias; semantic memory; computational models; social cognition

Introduction

Language is more than just a means of communication; it offers a window into a culture's concepts and social structures. Humans rely on their linguistic environment to derive shared word meanings within a given context. From an early age, statistical learning mechanisms help individuals infer structured relationships in their environments (Saffran & Kirkham, 2018). These mechanisms may also shape the organization of semantic knowledge, as captured by the distributional hypothesis – the idea that a word's meaning emerges from statistical patterns of its co-occurrence with other words in natural language (Firth, 1957; Harris, 1954).

Distributional semantic models (DSMs) operationalize this hypothesis by learning semantic relationships from large language corpora. These models represent word meanings as numerical vectors in a high-dimensional semantic space, where the distance between two word vectors (cosine similarity) reflects their semantic association. By leveraging co-occurrence patterns, DSMs have proven remarkably effective in modeling human semantic representations (Kumar, 2021).

In addition to capturing general conceptual associations, DSMs trained on large text corpora have been shown to encode well-documented implicit stereotypes, which are automatic associations between social categories and other content (Bolukbasi et al., 2016; Lauscher & Glavaš, 2019; Xu et al., 2019). One of the first demonstrations of this phenomenon came from Caliskan et al. (2017), who directly measured semantic biases in DSMs against implicit biases measured in previous social psychology experiments. They devised a test called the Word Embedding Association Test (WEAT), designed as a computational analogue to the Implicit Association Test (IAT).

The IAT measures association strength through differences in response latencies when categorizing stimuli (Greenwald et al., 1998). For example, in a gender-career IAT, "congruent" trials would involve using the one key for *men* and *career* words and another for *women* and *family* words. "Incongruent" trials would have the opposite pairing. The IAT score reflects the difference in response times when categorizing stimuli during congruent versus incongruent trials. The effect size of this response time differential is interpreted as a measure of implicit association strength. Similarly, the WEAT would quantify this association in DSMs by comparing the average cosine similarity of *men* and *career* versus *family* words and repeating this process for *women*. The final score is calculated by subtracting the difference for *women* from the difference for *men*. Caliskan et al. tested the WEAT on pre-trained GloVe embeddings (a widely used DSM; Pennington et al., 2014) with stimuli from previous behavioral implicit bias experiments and successfully replicated each of the effects. These findings showed that implicit social biases are embedded within the distributional structure of language as modeled by DSMs.

First- and Second-Order Association

Although these studies demonstrate that DSMs can encode human social biases, the specific origins of these biases within the models remain unclear. GloVe representations are constructed by combining of two types of distributional statistics: first-order association and second-order association (Levy et al., 2015).

First-order co-occurrence, also known as syntagmatic association, arises when two words frequently co-occur in the same linguistic context. For example, *boat* and *dock* are more

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

likely to occur together than *boat* and *eraser*. As a result, the former pair exhibits a stronger first-order association than the latter. This type of co-occurrence is closely linked to the concept of relatedness, reflecting how words are contextually associated. Second-order co-occurrence, also known as paradigmatic association, arises when two words frequently occur with the same context words but not with each other. Although *boat* and *ship* may not often directly co-occur, they frequently appear alongside the same sets of words as each other, establishing a second-order association between them. Second-order co-occurrence is widely regarded as a key mechanism underlying semantic similarity (Landauer & Dumais, 1997).

In addition to forming the basis for DSM algorithms, first- and second-order association play important roles in human cognition. First-order co-occurrence predicts performance on free association tasks while second-order co-occurrence predicts performance on synonymy generation tasks (Rapp, 2002). Both sources are used independently in learning word relations in statistical learning experiments (Unger et al., 2020; Yu & Smith, 2007), and children are known to display a changeover in preference for first- to second-order relations around the time of preschool (Unger et al., 2016). The differentiation in these statistical sources are the basis of mechanistic theories of the hippocampus and neocortex as complimentary learning systems that have evolved to capitalize on each source in the learning environment (McClelland et al., 1995).

Algorithms like GloVe measure word similarity by adding two types of information: context vectors, which increase similarity when words frequently co-occur in the same contexts (first-order co-occurrence), and word vectors, which increase similarity for words that are interchangeable or substitutable (second-order co-occurrence). Thus, the semantic biases found by Caliskan et al. (2017) may have been encoded through first-order statistics, second-order statistics, or a combination of both. To investigate the extent to which implicit biases are encoded by first- or second-order association, we replicated the original WEATs, isolating each type of co-occurrence to determine whether similar effects could be independently observed.

The Present Study

To examine how a DSM encodes stereotypes from a standard natural language corpus, we sought to determine whether these biases could be attributed to direct or indirect co-occurrence statistics. Previous research has found human implicit biases in first-order co-occurrence relations (Lynott et al., 2012; Rekabsaz et al., 2021; Valentini et al., 2023). However, no studies have directly compared first-order and second-order associations to those found in DSMs or explored how these associations vary across topics. In this study, we used the original WEAT stimuli and adapted the WEAT algorithm, replacing DSM cosine similarity measures with first- and second-order similarity metrics.

We also aimed to uncover the patterns driving stereotypic associations in first- and second-order statistics. Like the IAT, the WEAT condenses relative association scores between four categories into a single value, potentially obscuring nuanced patterns in the data. For instance, Bailey et al. (2022) found that gender bias in DSM embeddings is asymmetrical. While the words for *women* were more associated with stereotypically female than male traits, words for *men* were equally associated with all traits describing people. The WEAT alone would be unable to capture this kind of asymmetry. Therefore, we deconstructed each WEAT score into its components to pinpoint the sources of the observed associations and analyze how these vary across topics and between first- and second-order data. This approach provides a more granular perspective on the mechanisms underlying stereotypic associations, offering insights otherwise hidden in aggregated scores.

Our work identifies the implicit stereotypes that can theoretically be captured by simple mechanisms of first- and second-order statistical learning. By isolating these potential sources, this study advances our understanding of implicit biases, their origins, and the statistical signals by which they may be transmitted within natural language.

Method

Co-Occurrence Measures

To quantify first- and second-order co-occurrence of word pairs, we employed two variants of pointwise mutual information (PMI): Positive PMI (PPMI; Church and Hanks, 1990) for first-order co-occurrence and Second-Order Co-occurrence PMI (SOC-PMI; Islam and Inkpen, 2006) for second-order co-occurrence.

First-Order Co-Occurrence First-order association reflects the direct association between two words, measured by how frequently they co-occur within the same linguistic context. PPMI is an information-theoretic measure of the frequency with which two words w_1 and w_2 occur together relative to their base rates:

$$\mathrm{PPMI}(w_1, w_2) = \max\left(\log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}, 0\right). \quad (1)$$

$P(w_1, w_2)$ is the probability that w_1 and w_2 co-occur in the same context window. $P(w_1)$ and $P(w_2)$ are the probabilities of each word occurring independently. We use a window size of five words on each side of the target word (11 words total), which is the same setting used for the pre-trained GloVe model used by Caliskan et al. Since negative PMI values often reflect noise rather than meaningful relationships, PPMI replaces these scores with zero to improve reliability as a measure of relatedness.

Second-Order Co-Occurrence Second-order association emerges when two words co-occur with similar words as each other. SOC-PMI takes the highest-PMI neighbors of each word and calculates a similarity score from a weighted sum of the overlapping neighbors.

Following the algorithm outlined by Islam and Inkpen (2006, 2008), for both words w_1 and w_2, we define a set X^w of the words with the top β PMI values (greater than zero) with the word w. As in the first-order measure, we use a context window size of five words on each side of the target word. β is a function of the frequency of w, the number of types in the corpus, and a parameter set based on the size of the corpus. This value represents the number of first-order associates that will be considered and compared for each of the two target words, aiming to capture the most important neighbors of each. The PMI values of words in X^{w_1} that are also in X^{w_2} are summed using the following function, reflecting the contribution of overlapping neighbors to second-order association:

$$f(w_1, w_2, \beta) = \sum_{i=1}^{\beta} \left(\text{PMI}(\mathbf{X}_i^{w_1}, w_2)\right)^{\gamma}. \quad (2)$$

γ represents the emphasis placed on the words' PMI values with w, such that higher γ values will give greater weight to words with high PMI values. The values of γ and the corpus size parameter in β were hand-fit to differentiate between words from established similarity (Hill et al., 2015) and relatedness norms (Jouravlev & McRae, 2016) in our corpus. The final similarity score is calculated as:

$$\text{SOC-PMI}(w_1, w_2) = \frac{f(w_1, w_2, \beta_1)}{\beta_1} + \frac{f(w_2, w_1, \beta_2)}{\beta_2}. \quad (3)$$

Materials

Corpus To compare our results with those reported by Caliskan et al. (2017) using pre-trained GloVe embeddings, we selected a corpus comparable to the Common Crawl corpus used to train GloVe: the Colossal Cleaned Common Crawl (C4; Raffel et al., 2020). We calculated all first- and second-order co-occurrence measures using this corpus. The C4 corpus is a cleaned subset of the Common Crawl web scrape corpus, designed to remove noisy data and improve text quality for research applications. We used the validation set of the English-language version of the dataset, sourced from Hugging Face, to ensure a manageable and representative subset for analysis. After pre-processing the data using the spaCy library – lemmatizing tokens and removing stop words, punctuation, and non-alphabetic tokens – to reduce noise and standardize the text, the resulting corpus contained approximately 67.2 million tokens. This size was sufficient to capture robust co-occurrence patterns while remaining computationally manageable.

Stimuli We used the same stimuli as the original WEATs, which consist of predefined sets of target and attribute words. These stimuli are available in the Supplemental Materials of Caliskan et al. (2017). The ten tests each include two target category sets and two attribute category sets, which were originally collected from behavioral implicit association experiments. All the original WEATs performed on GloVe embeddings produced effects consistent with the experimental results, validating the method's ability to replicate human implicit biases.

While most of the stimuli are drawn from IAT studies, two tests specifically use common Black and White American names from research on implicit racial bias in hiring decisions. Following the methodology of Caliskan et al., we began with the full word sets for each test. If a term was missing from the corpus, we removed that word along with a randomly selected counterpart from the other target or attribute set. The ten tests cover a range of topics, including two non-social topics (flowers/insects and instruments/weapons), three variations of racial bias stimuli, three gender stereotypes, mental illness stigma, and age bias.

Analyses

Bias Effect Size To measure biases in first- and second-order language statistics, we modified the WEAT effect size calculation by replacing cosine similarity with either the first-order similarity score (PPMI) or second-order similarity score (SOC-PMI). Drawing from the terminology of the IAT, the WEAT compares two sets of target words of equal size, X and Y (e.g., *men* and *women*), and two sets of attribute words, A and B (e.g., *career* and *family*). The goal is of the WEAT is to quantify the relative association strength of X and A compared to Y and B (e.g., *men-career* versus *women-family*). The effect size is:

$$\frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std\_dev}_{w \in X \cup Y} s(w, A, B)} \quad (4)$$

where

$$s(w, A, B) = \text{mean}_{a \in A} \text{sim}(w, a) - \text{mean}_{b \in B} \text{sim}(w, b) \quad (5)$$

and $\text{sim}(w_1, w_2)$ denotes the co-occurrence function, which calculates either PPMI or SOC-PMI depending on the analysis.

The result is a normalized measure of X's average association with A relative to B plus Y's average association with B relative to A. As in Caliskan et al. (2017), we obtained one-sided p values for each effect size using a permutation test.

Disaggregating Bias Effects Within the WEAT framework, the effect size metric combines the association strengths of both X and Y with A and B. To better understand the specific language patterns contributing to these results, we disaggregated the similarity scores for each target category. For each social bias test, we broke down the effect sizes by calculating each target word's average association with the words in set A and those in set B. The WEAT calculates the difference between the average difference of these values for X and Y. By stopping before these steps and analyzing the constituent parts of the WEAT effect size, we can examine whether some of these difference scores are driving the overall score more than others. This disaggregation provides finer-grained insights into the contributions of individual targets to the overall bias effect. We performed independent samples t-tests to evaluate two key aspects: the differences between each target set's associations with A and B and

differences in associations between the two target sets. These comparisons highlight potential asymmetries in how targets are linked to attributes.

Results

Overall Results

As shown in Table 1, nine out of ten of the associations identified by Caliskan et al. (2017) were significant in first-order statistics ($M = 1.02$, $SD = 0.37$), while five were significant in second-order statistics ($M = 0.83$, $SD = 0.58$). Notably, each tested bias was captured in at least one type of co-occurrence information, suggesting that such biases can be accounted for by simple distributional mechanisms. For the two non-social topics (rows 1 and 2), the associations of interest were robust across both first- and second-order co-occurrences. Among the social topics, first-order associations consistently produced stronger effects than second-order associations, with the exception of mental versus physical illness (row 9), where second-order effects were larger. The following sections deconstruct these effect sizes for the social topics.

By Topic

Race Of the racial bias tests (Table 1, rows 3-5), all demonstrated significant first-order effects, but only one (row 3) had an equally strong second-order effect. The tests in rows 3 and 4 used the same attribute words but different names, and the tests in rows 4 and 5 used the same names but different attribute words. Because the three tests showed qualitatively similar results when broken down by category, Figure 1a and e show the disaggregated association scores averaged across the three tests for the sake of space. The results for each individual test are described in detail below.

For row 3, both White names ($t(76) = 7.29,\ p < .001$) and Black names ($t(76) = 2.42,\ p = .018$) showed higher mean PPMI scores with *pleasant* than with *unpleasant*. However, both attributes' scores were higher for White names than Black names, $t(76) = 17.20,\ p < .001$ for *pleasant* and $t(76) = 12.50,\ p < .001$ for *unpleasant* words. The overall effect was because the *pleasant-unpleasant* association gap was wider for White than Black names. In contrast, second-order scores revealed no difference between *pleasant* and *unpleasant* associations for White names ($t(76) = -0.35,\ p = .728$), while Black names showed a stronger association with *unpleasant* than *pleasant* words ($t(76) = 3.79,\ p < .001$). The second-order bias was driven by Black names' heightened association with *unpleasant* words.

For row 4, the first-order pattern mirrored that of row 3, but the difference between *pleasant* and *unpleasant* associations for Black names was not significant, $t(34) = 1.19,\ p = .243$. In second-order scores, both White and Black names had equivalent similarity to *pleasant* words ($t(34) = -1.22,\ p = .229$), but Black names were significantly more associated with *unpleasant* words ($t(34) = 3.49,\ p = .001$). The second-order effect size was lower than in row 3, as White names also showed a higher association with *unpleasant* than *pleasant* words, albeit to a lesser extent, $t(34) = 2.08,\ p = .045$.

For row 5, the first-order pattern was consistent with row 3. However, second-order scores revealed no significant differences between White and Black names in their associations with either *pleasant* ($t(34) = -0.30,\ p = .764$) or *unpleasant* words ($t(34) = 0.34,\ p = .735$). Consequently, no overall bias was evident in second-order associations for this test. Because it used the same name stimuli as row 4, which showed a moderate second-order effect, the lack of effect for this test is a result of the sets of attribute words used, either because of their smaller sizes or the specific words included.

Gender The gender bias tests (rows 6-8) include one test where gender categories served as targets and *career* and *family* words as attributes (row 6). The other two tests examined disciplines as targets and gender categories as attributes (rows 7-8). While all three tests revealed significant first-order effects, only the gender-career test exhibited a strong second-order effect.

The gender-career stereotype showed large effect sizes driven by the co-occurrence patterns of women's names. In first-order statistics (Figure 1b), men's names were equally associated with *career* and *family* words, $t(14) = 0.12,\ p = .910$. In contrast, women's *career* and *family* scores differed significantly, $t(14) = -6.31,\ p < .001$. Women's names were less strongly associated with *career* than men's ($t(14) = -5.34,\ p < .001$) and more strongly associated with *family* ($t(14) = 3.22,\ p = .006$). In second-order associations (Figure 1f), a similar pattern emerged. Women's names showed an equal association to *career* as men's ($t(14) = -0.23,\ p = .821$), but their association with *family* remained significantly higher ($t(14) = 4.76,\ p < .001$).

Although the gender differences in the *math* and *arts* associations were not individually significant, they followed the expected pattern (i.e, *math* had a stronger association with *men*; *arts* had a stronger association with *women*). When combined, their difference scores produced a significant overall first-order effect. *Arts* was more strongly associated than *math* with both *men* ($t(14) = 2.96,\ p = .010$) and *women* ($t(14) = 3.41,\ p = .004$). This pattern held for second-order associations, though the overall effect size was notably smaller.

The *science* versus *arts* test showed results similar to those observed in the *math* and *arts* test. However, the first-order associations between *science* and *men* versus *women* differed significantly, $t(14) = 2.61,\ p = .020$. The second-order scores showed no significant differences from each other.

Mental Illness Unlike other stereotypes tested, the mental illness stereotype (row 9) exhibited no significant first-order bias but showed a pronounced second-order bias. Further analysis revealed that both mental and physical illness words were more likely to directly co-occur with *permanent* than *temporary*, $t(10) = 2.73,\ p = .021$ and $t(10) = 3.13,\ p = .011$, respectively (Figure 1c). Their scores for each attribute

Table 1: Summary of results, including the number of words per set in the pair (N).

Target words (N)	Attribute words (N)	First-Order		Second-Order	
		d	p	d	p
Flowers vs. insects (25)	Pleasant vs. unpleasant (25)	**1.02**	$< .001$	**1.39**	$< .001$
Instruments vs. weapons (25)	Pleasant vs. unpleasant (25)	**1.37**	$< .001$	**1.31**	$< .001$
White vs. Black American names (39)	Pleasant vs. unpleasant (25)	**0.92**	$< .001$	**0.91**	$< .001$
White vs. Black American names (18)	Pleasant vs. unpleasant (25)	**0.82**	0.003	0.54	0.059
White vs. Black American names (18)	Pleasant vs. unpleasant (8)	**0.63**	0.024	-0.18	0.705
Male vs. female names (8)	Career vs. family (8)	**1.67**	$< .001$	**1.40**	0.002
Math vs. arts (8)	Male vs. female terms (7)	**1.20**	0.008	0.45	0.191
Science vs. arts (8)	Male vs. female terms (7)	**1.33**	0.001	0.25	0.306
Mental vs. physical illness (6)	Temporary vs. permanent (7)	0.34	0.268	**1.71**	$< .001$
Young vs. old people's names (8)	Pleasant vs. unpleasant (8)	**0.88**	0.047	0.56	0.146

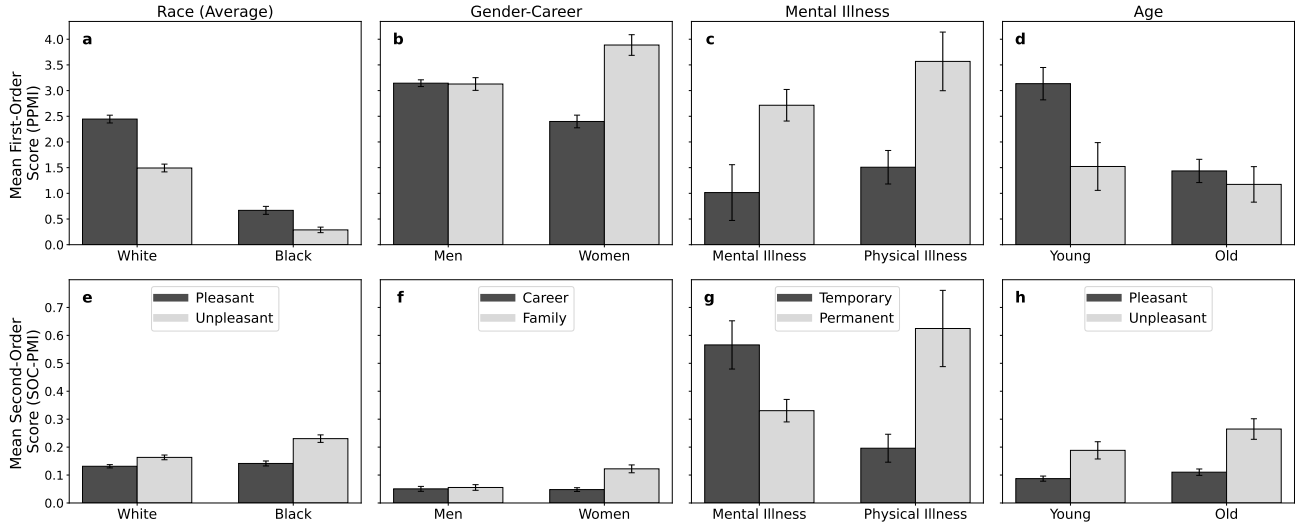

Figure 1: Average association score of each target category to each attribute category in first-order (a-d) and second-order co-occurrence (e-h) for selected tests, representing each of the main social topics. Overall effect sizes are based on a double difference score of these values: the difference between each target's difference in score for the two attributes. Error bars show standard errors.

were not significantly different from each other. However, in second-order association, the relationship flipped for mental illness words (Figure 1g). *Mental illness* was significantly more associated with *temporary* words than *permanent* ones ($t(10) = 2.47,\ p = .033$), while *physical illness* remained more associated to *permanent* words than *temporary* ones ($t(10) = 2.95,\ p = .015$). The relative association between mental illness and impermanence emerged exclusively in second-order co-occurrences, sharply contrasting with the co-occurrence patterns for physical illness words.

Age Age bias (row 10) exhibited a stronger first-order than second-order effect, as measured by associations between typically younger and older names and *pleasant* or *unpleasant* words. Both young and old names were more directly associated with *pleasant* than *unpleasant* (Figure 1d). However, this difference was significant for young names ($t(14) = 2.87,\ p = .012$) and non-significant for old names ($t(14) = 0.63,\ p = .537$). The overall first-order effect was primarily driven by the greater co-occurrence for younger names with *pleasant* words compared to older names. In second-order associations, both young and old names were more strongly associated with *unpleasant* than *pleasant* (Figure 1h). This difference was larger for old names ($t(14) = 4.01,\ p = .001$) than for young names ($t(14) = 3.14,\ p = .007$).

Discussion

To investigate how DSMs form social biases from natural language, we isolated two types of co-occurrence information used to construct their embeddings: first- and second-order associations. First-order association arises when words frequently appear nearby each other, while second-order association reflects words that have similar first-order associates as each other. By measuring the effect sizes of several previ-

ously studied biases using each type of co-occurrence, we assessed the extent to which each source could contribute to the biases learned by DSMs. Our findings indicate that these simple distributional statistics are sufficient to account for many of the social biases observed in more complex semantic models.

The vast majority of the implicit associations were present in first-order statistics, showing that they can be captured by merely tracking direct co-occurrence patterns. While the non-social topics showed strong effects in both first- and second-order association, most social biases were more prominent in first-order than in second-order measures. This contrast may shed light on how the distributional structures of language differentially encode social versus non-social semantic associations.

The results, when analyzed in detail, reveal complex variation in how different stereotypes manifest within natural language statistics. For example, in contrast with the other topics, the mental illness-related stereotype had a small first-order effect but a very large second-order effect. The analyses reveal that the association between *mental illness* and *temporary*, as opposed to *permanent*, emerges only in second-order co-occurrences. This pattern suggests that stereotypes about mental illness may be expressed more indirectly in language, perhaps because explicitly describing mental illness in unstable terms is less socially acceptable.

We also found that each of the categories measured in relation to pleasant and unpleasant words (racial and age groups) had stronger associations with *pleasant* words in first-order statistics and with *unpleasant* words in second-order statistics. However, the magnitude of these differences was greater for the non-marginalized groups in first-order statistics and for the marginalized groups in second-order statistics. This pattern may suggest that people tend to use more direct language when describing others positively but rely on more indirect or subtle language when expressing negativity – especially with regard to marginalized groups.

Meanwhile, the gender-career association was entirely driven by the usage of women's names. Men's names showed equal co-occurrence with *career* and *family* words, while women's names were less associated with *career* words and more associated with *family* words. This result is in line with previous findings from Bailey et al. (2022) that the concept of *women* is semantically closer to stereotypically feminine than masculine traits while the concept *men* shows no such difference. Future research should investigate how different hegemonic systems create biases stemming from associations with the dominant group, with the marginalized group, or with both equally.

The work of researchers like Caliskan et al. (2017) proposed a parsimonious theoretical explanation for implicit biases by showing that they can be learned by computational models from the statistical regularities of natural language. The present study goes a step further in showing that the same biases can be captured by the simpler mechanisms that underlie these models, as well as parts of human cognition. By uncovering the specific language patterns that contribute to stereotypes in DSMs, we demystify how implicit biases are embedded in their training data and, possibly, our broader environment.

The finding that well-documented implicit associations mirror basic word co-occurrence patterns supports context-centric theories of implicit bias, such as the Bias of Crowds theory (Payne & Hannay, 2021). Concept accessibility in memory reflects statistical regularities in the environment, making a concept more accessible when it is cued by a frequent collocate. The Bias of Crowds theory suggests that the same mechanism underlies implicit biases measured by tests like the IAT. According to this view, structural inequalities in the social environment create biased co-occurrence patterns between social groups and stereotypic content. These patterns, in turn, make stereotypic associations more automatically accessible, on average. Our data, based on a large sample of English text from the Internet, supports the premise that implicit stereotypes are present in the low-level statistical regularities of the American cultural environment.

It is possible that the near ubiquity of first-order bias we observed only applies to our web crawl data and does not generalize to all language environments. Previous studies have found implicit biases in DSMs trained on various data sources, such as news articles, books, and child-directed speech (Bhatia, 2017; Charlesworth et al., 2021; Garg et al., 2018; Lewis et al., 2022). Future work should examine how these first- and second-order effects manifest across different linguistic domains. In other contexts where DSMs capture social biases, variation in situational norms will likely influence whether the biases arise from first- or second-order effects, or both.

The present research also has important implications for the ethical development of artificial intelligence (AI). Most modern language models rely on distributional statistics to construct their semantic representations. The present study demonstrates how these systems can internalize harmful biases from natural language data. Moreover, IAT scores are weak predictors of individual human behavior (Oswald et al., 2013, 2015). When people have sufficient cognitive resources, they typically avoid making decisions based solely on immediately evoked semantic associations. However, natural language processing systems are designed to replicate the relationships in their training data. As a result, they are predisposed to perpetuate implicit stereotypes, with far-reaching consequences as they become more integrated into society.

Our results demonstrate that implicit stereotypes are embedded in the statistical regularities of natural language – specifically, in first- and second-order distributional statistics. This suggests that any system sensitive to such patterns, whether a DSM or a human mind, can acquire biased associations from the input of its cultural environment, even without holding biased attitudes or beliefs.

Acknowledgments

This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

References

Bailey, A. H., Williams, A., & Cimpian, A. (2022). Based on billions of words on the internet, people= men. *Science Advances*, *8*(13), eabm2463.

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

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. *Advances in neural information processing systems*, *29*.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Charlesworth, T. E., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, *32*(2), 218–240.

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, *16*(1), 22–29.

Firth, J. (1957). A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis, Special Volume/Blackwell*.

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.

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.

Harris, Z. S. (1954). Distributional structure.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695.

Islam, A., & Inkpen, D. (2006). Second order co-occurrence pmi for determining the semantic similarity of words. *LREC*, 1033–1038.

Islam, A., & Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *2*(2), 1–25.

Jouravlev, O., & McRae, K. (2016). Thematic relatedness production norms for 100 object concepts. *Behavior research methods*, *48*, 1349–1357.

Kumar, A. A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, *28*(1), 40–80.

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.

Lauscher, A., & Glavaš, G. (2019). Are we consistently biased? multidimensional analysis of biases in distributional word vectors. *NAACL HLT 2019*, 85.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the association for computational linguistics*, *3*, 211–225.

Lewis, M., Cooper Borkenhagen, M., Converse, E., Lupyan, G., & Seidenberg, M. S. (2022). What might books be teaching young children about gender? *Psychological science*, *33*(1), 33–47.

Lynott, D., Kansal, H., Connell, L., & O'Brien, K. (2012). Modelling the iat: Implicit association test reflects shallow linguistic environment and not deep personal attitudes. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*(34).

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, *102*(3), 419.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of iat criterion studies. *Journal of personality and social psychology*, *105*(2), 171.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2015). Using the iat to predict ethnic and racial discrimination: Small effect sizes of unknown societal significance. *Journal of personality and social psychology*.

Payne, B. K., & Hannay, J. W. (2021). Implicit bias reflects systemic racism. *Trends in cognitive sciences*, *25*(11), 927–936.

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

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, *21*(140), 1–67.

Rapp, R. (2002). The computation of word associations: Comparing syntagmatic and paradigmatic approaches. *COLING 2002: The 19th International Conference on Computational Linguistics*.

Rekabsaz, N., West, R., Henderson, J., & Hanbury, A. (2021). Measuring societal biases from text corpora with smoothed first-order co-occurrence. *Proceedings of the international aaai conference on web and social media*, *15*, 549–560.

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual review of psychology*, *69*(1), 181–203.

Unger, L., Fisher, A. V., Nugent, R., Ventura, S. L., & MacLellan, C. J. (2016). Developmental changes in seman-

tic knowledge organization. *Journal of experimental child psychology*, *146*, 202–222.

Unger, L., Vales, C., & Fisher, A. V. (2020). The role of co-occurrence statistics in developing semantic knowledge. *Cognitive Science*, *44*(9), e12894.

Valentini, F., Rosati, G., Blasi, D., Slezak, D. F., & Altszyler, E. (2023). On the interpretability and significance of bias metrics in texts: A pmi-based approach. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 509–520.

Xu, H., Zhang, Z., Wu, L., & Wang, C.-J. (2019). The cinderella complex: Word embeddings reveal gender stereotypes in movies and books. *PloS one*, *14*(11), e0225385.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological science*, *18*(5), 414–420.