Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024)

ML-EAT: A Multilevel Embedding Association Test for Interpretable and Transparent Social Science

Robert Wolfe, Alexis Hiniker, Bill Howe

University of Washington
rwolfe3@uw.edu, alexisr@uw.edu, billhowe@uw.edu

Abstract

This research introduces the Multilevel Embedding Association Test (ML-EAT), a method designed for interpretable and transparent measurement of intrinsic bias in language technologies. The ML-EAT addresses issues of ambiguity and difficulty in interpreting the traditional EAT measurement by quantifying bias at three levels of increasing granularity: the differential association between two target concepts with two attribute concepts; the individual effect size of each target concept with two attribute concepts; and the association between each individual target concept and each individual attribute concept. Using the ML-EAT, this research defines a taxonomy of EAT patterns describing the nine possible outcomes of an embedding association test, each of which is associated with a unique EAT-Map, a novel four-quadrant visualization for interpreting the ML-EAT. Empirical analysis of static and diachronic word embeddings, GPT-2 language models, and a CLIP language-and-image model shows that EAT patterns add otherwise unobservable information about the component biases that make up an EAT; reveal the effects of prompting in zero-shot models; and can also identify situations when cosine similarity is an ineffective metric, rendering an EAT unreliable. Our work contributes a method for rendering bias more observable and interpretable, improving the transparency of computational investigations into human minds and societies.

Introduction

Computational methods that quantify societal biases using language technologies like word embeddings (Mikolov et al. 2013), generative language models (Radford et al. 2018), and multimodal language-and-image models (Radford et al. 2021) have been widely adopted by social scientists, who leverage the reflection of society captured by these technologies to observe implicit and explicit societal biases where human-subjects experiments are infeasible or prohibitively expensive (Durrheim et al. 2023; Kennedy et al. 2021). Social scientists have employed word embeddings in particular to analyze diachronic historical changes in human biases and norms (Garg et al. 2018), to study variations in gender biases across numerous languages (Lewis and Lupyan 2020), to compare implicit biases in adult and children's language corpora (Charlesworth et al. 2021), and to validate longstanding theories about societal biases, such as the masculine default (Caliskan et al. 2022; Bailey, Williams, and Cimpian 2022). Bhatia and Walasek (2023) employ such techniques to *predict* human biases, potentially facilitating mitigations at the societal scale. Among the most widely adopted bias measurement methods in computational social science is the Word Embedding Association Test (WEAT) (Caliskan, Bryson, and Narayanan 2017), a statistical technique grounded in the Implicit Association Test (IAT), a widely used measurement of unconscious bias in human subjects (Greenwald, McGhee, and Schwartz 1998). The WEAT quantifies bias based on the differential cosine similarity of two target groups X and Y (such as Science vs. Art) with two attribute groups A and B (such as Male vs. Female, for a common test of gender bias).

Yet the WEAT also suffers from a limitation common in AI evaluation: it is an aggregate metric (Burnell et al. 2023), averaging over many sub-measurements between groups of words to produce a single summary statistic. Explaining a bias quantified by the WEAT is thus not straightforward. In plain English, a statistically significant WEAT indicates that the X target group is more associated with A relative to B than the Y target group is more associated with A relative to B. While effective for surfacing implicit biases, the method also raises questions about whether and to what extent each target group is *individually* associated with A or B. Consider the test of age bias presented by Caliskan, Bryson, and Narayanan (2017), which sets group A to Pleasantness, B to Unpleasantness, X to Young Names, and Y to Old Names and returns a large, statistically significant effect size of 1.21. While one might intuitively interpret the result of the test to mean that Young Names are associated with Pleasantness, and Old Names with Unpleasantness, examining component cosine similarities used to compute the WEAT reveals that *both* Young Names (X) and Old Names (Y) are associated with Pleasantness (A). On the other hand, consider the WEAT setting A to Pleasantness, B to Unpleasantness, X to Instruments, and Y to Weapons, which also returns a large, significant effect size, of 1.53. Inspection of component cosine similarities reveals that, in this case, Instruments are associated with Pleasantness, while Weapons are associated with Unpleasantness. Though the two WEATs return effect sizes with the same sign and similarly large magnitudes, the characteristics of their underlying biases, and the corresponding sociological interpretations of the results, differ significantly. Given the wide range of psychological and computational

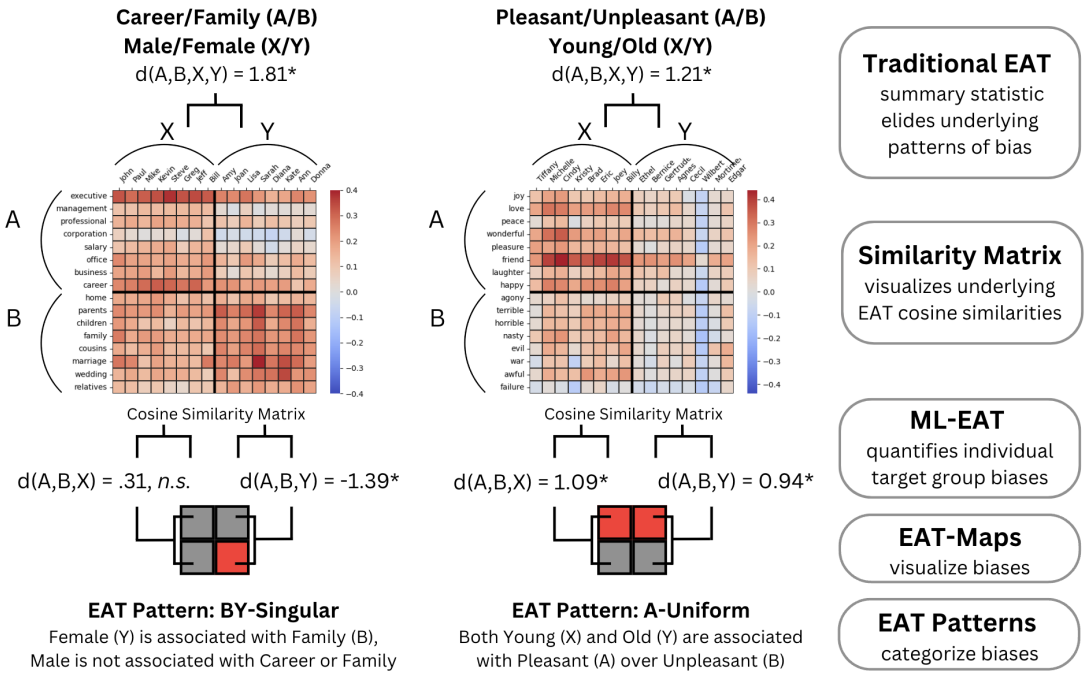

Figure 1: A visualization of the ML-EAT applied to the Career/Family and Young/Old EATs introduced by Caliskan, Bryson, and Narayanan (2017). Where traditional EATs return a single effect size and p-value, the ML-EAT surfaces underlying patterns of bias in individual target group associations with the A or B attribute.

studies in which WEAT scores are utilized, knowing the difference between these outcomes can be important - but they are indistinguishable given a single WEAT d-score.

The present work provides a method and a vocabulary for describing meaningful differences in the underlying biases quantified by EATs (Embedding Association Tests, which are now widely utilized beyond word embeddings (Steed and Caliskan 2021)). We make the following contributions:

1. The **MultiLevel Embedding Association Test (ML-EAT)**, a formal, theoretically justified measure to disambiguate potential sources of bias in machine learned representations. The ML-EAT quantifies bias at three levels: Level 1, an effect size describing the differential association between two target concepts with two attribute concepts (equivalent to a traditional WEAT); Level 2, two effect sizes describing the differential association of each target group X and Y *individually* with the two attribute groups A and B; and Level 3, the means and standard deviations of the four underlying distributions of cosine similarities (X, A), (X, B), (Y, A), (Y, B) that ultimately compose an EAT measurement.
2. **A taxonomy of EAT patterns** for characterizing the nine discrete outcomes of EATs, each of which admits a distinct interpretation. The taxonomy provides a vocabulary for the underlying associations that make up an EAT, based on whether each of the target groups X and Y individually exhibits a significant bias toward A or B.
3. **The EAT-Map, an intuitive visualization for interpreting the ML-EAT**. As illustrated near the bottom of Figure 1, the EAT-Map uses a four-quadrant square with columns corresponding to the target groups X and Y and rows corresponding to the attribute groups A and B, and shades cells in order to denote associations between a target group and an attribute. The EAT-Map provides a visual vocabulary for the taxonomy of EAT patterns, as each EAT pattern has a unique EAT-Map.
4. **An empirical analysis of the ML-EAT** applied to static and diachronic word embeddings (Pennington, Socher, and Manning 2014; Hamilton, Leskovec, and Jurafsky 2016), GPT-2 generative language models (Radford et al. 2019), and a CLIP language-and-image model (Radford et al. 2021). Applying the ML-EAT to the GloVe embeddings shows that five distinct EAT patterns occur in the ten WEATs performed by Caliskan, Bryson, and Narayanan (2017), while using the test with the HistWords embeddings (Hamilton, Leskovec, and Jurafsky 2016) shows that EAT patterns help to draw more complete conclusions about historical biases. Analysis of GPT-2 and CLIP models further demonstrates that the ML-EAT can surface the effects of prompting in the zero-shot setting, as well as identify embedding spaces unsuitable for analysis with cosine similarity.

The ML-EAT provides an expressive means to describe bias in language technologies. As we will demonstrate, even well-studied results, such as biases measured by Caliskan, Bryson, and Narayanan (2017), yield new perspectives with the ML-EAT. Our research code is available at https://github.com/wolferobert3/ml-eat.

Related Work

In reviewing the related on work on Embedding Association Tests (EATs), we describe the models and domains in which EATs are employed; then provide a detailed overview of their applications in computational social science; and finally review their limitations as described in prior work.

Embedding Association Tests

Caliskan, Bryson, and Narayanan (2017) introduce the Word Embedding Association Test (WEAT), a measurement of intrinsic bias in word embeddings drawing on the design of the Implicit Association Test (IAT), a method for studying implicit (unconscious) bias in human subjects (Greenwald, McGhee, and Schwartz 1998). The WEAT quantifies the relative association of two target groups (such as Science and Art) with two attribute groups (such as Male and Female), and, like the IAT, returns an effect size (Cohen's d) and a p-value. Caliskan, Bryson, and Narayanan (2017) used the WEAT to replicate the results of ten IATs reflecting gender and racial biases, among others, in the GloVe embeddings of Pennington, Socher, and Manning (2014). Caliskan, Bryson, and Narayanan (2017) also introduced the Single Category SC-WEAT, which quantifies the differential association of a single word with two attribute groups. The SC-WEAT was introduced as part of a measurement called the Word Embedding Factual Association Test, and further clarified in an analysis of androcentric gender bias by Caliskan et al. (2022).

The WEAT was extended to transformer models by May et al. (2019), who introduced the Sentence Embedding Association Test (SEAT), a sentence-level WEAT employing semantically neutral prompts, and by Kurita et al. (2019), who tied bias measurement to the model's pretraining objective. Guo and Caliskan (2021) introduced the Contextualized Embedding Association Test (CEAT), which measured embedding bias at the word level and modeled contextualization as a random effect. EATs are also used in computer vision and multimodal language-and-image models. Steed and Caliskan (2021) introduced the Image Embedding Association Test (iEAT), which quantified bias in self-supervised image encoders such as SimCLR (Chen et al. 2020b) and iGPT (Chen et al. 2020a). Ross, Katz, and Barbu (2020) introduced the Grounded-WEAT for grounded language-and-image models, while (Wolfe and Caliskan 2022a) use an EAT to quantify bias in CLIP language-and-image models (Radford et al. 2021), and Hausladen et al. (2023) employ the SC-EAT to perform a causal analysis of social bias in language-and-image models. Finally, Slaughter et al. (2023) introduce the SpEAT, an EAT for quantifying intrinsic bias in pretained speech processing models such as wav2vec2 (Baevski et al. 2020) and OpenAI Whisper (Radford et al. 2023).

While most EATs employ cosine similarity to measure association between target and attribute stimuli, recent work explores alternatives. Omrani Sabbaghi, Wolfe, and Caliskan (2023) employ an algebraic definition of bias similar to that of Bolukbasi et al. (2016), but use an EAT-based formula to obtain an effect size and p-value. Bai et al. (2024) assess implicit bias using the textual output of ostensibly debiased generative language models by employing a prompt-based analogue of the IAT.

Applications of EATs in Social Science

Social scientists use word embeddings and EATs to study phenomena that are impossible or financially infeasible to measure solely through direct experimental methods. Many studies leverage the societal scale of the data used for training word embeddings to make broad inferences about society that would be unavailable in a small-N psychological study. Caliskan et al. (2022) demonstrate a masculine default in the English-language internet using EATs computed for every word in the vocabulary of pretrained GloVe and FastText embeddings, while Bailey, Williams, and Cimpian (2022) use word embeddings to demonstrate the implicit equivalence of the concept of "person" with "men." Napp (2023) use EAT measurements to contend that gender stereotypes are stronger in countries that are more economically developed and individualistic. Finally, Schmahl et al. (2020) use the EAT to study changes in gender bias on Wikipedia, informing their suggestions for reducing bias on the platform. Other research employs word embeddings for previously untenable cross-cultural analyses of human attitudes. For example, Mukherjee et al. (2023) employ EAT to measure biases related to albeism, immigration, and education across 24 languages.

Embeddings of historical data have also provided a means to study human societies that no longer exist and are thus unavailable for direct study. Charlesworth, Caliskan, and Banaji (2022) use word embeddings to measure changes in stereotypes about social groups over 200 years. Borenstein et al. (2023) use the EAT to study intersectional biases in word embeddings trained on newspapers from the 18th and 19th century. Sunsay (2023) use the EAT to study the disease avoidance theory of xenophobia, measuring EAT scores in 19th and 20th century travel literature to test western associations of indigenous people disgust words that would suggest disease avoidance. Leach, Kitchin, and Sutton (2023) employ the EAT to argue that "language has changed in a way that reflects greater concern for others," supporting the idea that the societal "moral circle" expanded during the 19th and 20th centuries to include more groups of people, in addition animals and the environment. Guan et al. (2024) use cosine similarities to measure the evolution of color associations (*e.g.,* association of the color red with heat) over 200 years, while Betti, Abrate, and Kaltenbrunner (2023) use the EAT to measure sexism in fifty years of English song lyrics. Finally, Wolfe, Banaji, and Caliskan (2022) find evidence of the historical bias of hypodescent in CLIP models.

Research also employs EATs to study present-day bias in domains such as law and medicine. Rios, Joshi, and Shin (2020) use the WEAT to measure gender bias in biomedical research, finding that traditional gender stereotypes have declined over time, but that specific medical conditions like body dysmorphia still exhibit high gender bias. Cobert et al. (2024) use cosine similarities to measure implicit racial biases in ICU notes. In the legal domain, Matthews, Hudzina, and Sepehr (2022) use the EAT to study biases in word embeddings trained on corpora of legal opinions, and Dutta et al. (2023) use WEAT scores to quantify gender bias in Indian divorce court proceedings. Moreover, amid increasing interest in using AI to measure aspects of human society (Park et al. 2023; Shanahan, McDonell, and Reynolds 2023;

Xu et al. 2023), scholars have used EATs in an attempt to *predict* human attitudes. For example, Bhatia and Walasek (2023) use WEAT scores and a novel Valence Estimation Model (VEM) to predict human implicit biases. Similarly, Morehouse et al. (2023) measure the relationship of implicit and explicit human biases using the IAT, the WEAT, and the Mean Average Cosine method (Manzini et al. 2019).

Limitations of EATs

The EAT faces challenges related both to its mathematical definition and its predictive value in NLP applications. Social scientists who choose not to use the EAT sometimes note that its design, which mimics the differential construction of the IAT, can cause difficulties in observing and interpreting bias. Bailey, Williams, and Cimpian (2022) use the difference in raw cosine similarities to observe asymmetrical gender biases, noting that the WEAT is better applied to symmetrical patterns of association. Ethayarajh, Duvenaud, and Hirst (2019) contend that the standardization of the WEAT, in dividing by the joint standard deviation of word associations, can obscure differences in underlying cosine similarities. Moreover, anisotropy (directional uniformity) in deep learning models (Mu and Viswanath 2018) can distort intrinsic semantic measurements (Timkey and van Schijndel 2021), necessitating postprocessing of EATs (Wolfe and Caliskan 2022b).

Recent work suggests that EATs have limited predictive value for bias in downstream NLP tasks. Goldfarb-Tarrant et al. (2020) find that biases observed with the WEAT are not correlated with application biases in tasks like coreference resolution. In a study of downstream propagation using transformer language models, Orgad, Goldfarb-Tarrant, and Belinkov (2022) propose an information-theoretic framework rather than an EAT. Cabello, Jørgensen, and Søgaard (2023) show that association bias and fairness are uncorrelated, but also provide sociological evidence that the two kinds of metrics should be expected to be independent of each other.

We are concerned with uses of the EAT in social science to study human attitudes. To that end, we design an interpretable EAT, rather than an EAT predictive of downstream bias.

Models and Data

This research introduces the ML-EAT and applies it to word embeddings, GPT-2, and CLIP, adapting stimuli from prior studies of implicit bias in NLP and computer vision.

Pretrained Models

We apply the ML-EAT to the below language technologies.

- **GloVe Word Embeddings**: **Glo**bal **Ve**ctors for word representation (GloVe) train on the co-occurrence matrix of a text corpus, such that the vector representation of a word is learned based on the words it is most likely to occur around (Pennington, Socher, and Manning 2014). Caliskan, Bryson, and Narayanan (2017) introduced the WEAT by presenting results on 300-dimensional GloVe vectors trained on the 840-billion token Common Crawl.
- **HistWords Embeddings**: HistWords refers to sets of 20 word embeddings in four languages trained on ten-year slices of historical language corpora ranging between the years 1800 and 2000 (Hamilton, Leskovec, and Jurafsky 2016). Hamilton, Leskovec, and Jurafsky (2016) introduced HistWords to prove that more frequently used words exhibit less semantic change over time, and that polysemous words exhibit faster semantic change. We apply the ML-EAT to the English language HistWords embeddings trained using Word2Vec (SGNS) (Mikolov et al. 2013) on Google books (all genres) (Lin et al. 2012).
- **GPT-2 Language Models:** GPT-2 ("Generative Pretrained Transformer") is a causally masked transformer (Vaswani et al. 2017) language model trained to predict the next word in a sequence (Radford et al. 2019). This research studies the four pretrained GPT-2 models (Base, Medium, Large, and XL) available via the Transformers library (Wolf et al. 2020), which were pretrained on OpenAI's WebText dataset, a collection of webpages scraped from highly rated outbound links on Reddit.
- **CLIP Language-and-Image Models:** CLIP ("Contrastive Language Image Pretraining") is a multimodal language-and-image model, which classifies images based on their cosine similarity with text labels (Radford et al. 2021). This research reports results under varying prompts from the CLIP-ViT-L14-336 model, the best performing OpenAI-trained CLIP model available. CLIP-ViT-L14-336 is trained on OpenAI's WebImageText (WIT) dataset, a collection of 400 million pairs of web-scraped images and accompanying captions (Radford et al. 2021).

GPT-2 embeddings are obtained from the model's top layer, consistent with both May et al. (2019) and Guo and Caliskan (2021). CLIP embeddings are collected after projection to the model's multimodal text-and-image latent space.

EAT Stimuli

An EAT employs four groups of words or images (called "stimuli", drawing on the test's psychological foundations in the IAT (Greenwald, McGhee, and Schwartz 1998)), each representing a concept. For example, the EAT demonstrating that flowers are favored over insects uses word lists to represent the concepts of Flowers, Insects, Pleasant, and Unpleasant. Each EAT includes two "target" groups, X and Y (Flowers and Insects), which are tested for association with two "attribute" groups, A and B (Pleasant and Unpleasant) (Caliskan, Bryson, and Narayanan 2017). The two target groups contain the same number of stimuli, as do the two attribute groups. Groups must contain at least eight stimuli to adequately represent a concept (Caliskan et al. 2022).

We use the stimuli for the tests of implicit bias specified by Caliskan, Bryson, and Narayanan (2017) when applying the ML-EAT to GloVe and GPT-2. We applied the Math/Arts Male/Female EAT to the HistWords embeddings, replacing three stimuli because their L2 norms were zero-valued (preventing the computation of cosine similarity) in several HistWords embeddings. We substituted "music" for "symphony"; "mathematics" for "math"; and "calculation" for "calculus." Tests of bias in CLIP utilize the word stimuli of Caliskan, Bryson, and Narayanan (2017) to represent Pleasant and Unpleasant, and image stimuli from Steed and Caliskan (2021).

EAT Pattern	EAT-Map	Direction	Associations	WEAT Example
AB-Divergent		Divergent	X->A, Y->B	Flowers/Insects, P/U25
BA-Divergent		Divergent	X->B, Y->A	N/A
A-Uniform		Uniform	X->A, Y->A	Young/Old, P/U25
B-Uniform		Uniform	X->B, Y->B	N/A
AX-Singular		Singular	X->A	Science/Arts, Male/Female
BX-Singular		Singular	X->B	N/A
AY-Singular		Singular	Y->A	N/A
BY-Singular		Singular	Y->B	Male/Female, Career/Family
Non-Directional		None	None	Math/Arts, Male/Female

Figure 2: A taxonomy of EAT patterns describes associations of an EAT's target groups with its two attribute groups. Each pattern has a unique EAT-Map formed by shading cells of significant Level 2 tests, with target groups on the X-axis and attributes on the Y.

Approach

The ML-EAT is defined using three levels of measurement, with a taxonomy of nine EAT patterns for describing biases it quantifies. We first describe the test itself, then introduce the EAT-Map visualization and EAT pattern taxonomy.

Defining the ML-EAT

The ML-EAT computes bias at three levels of increasing granularity. Level 1 returns the traditional standardized effect size quantifying the differential association between two target concepts with two attribute concepts; Level 2 returns two effect sizes quantifying the differential association of each target group individually with the two attribute groups; and Level 3 returns four means and corresponding standard deviations describing the non-differential association of each target group and attribute group.

Level 1 The first level of the ML-EAT is equivalent to the WEAT, as given by Caliskan, Bryson, and Narayanan (2017):

$$\frac{\text{mean}_{x\in X} s(x, A, B) - \text{mean}_{y\in Y} s(y, A, B)}{\text{std\_dev}_{w\in X\cup Y} s(w, A, B)} \quad (1)$$

where A and B are attributes, X and Y are target groups, and association *s()* for an embedding $\vec{w}$ is:

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

This means that the association *s()* for each target stimulus represented by $\vec{w}$ is equal to its average association with the attribute stimuli in A, minus its average association with the attribute stimuli in B. The EAT returns an effect size (Cohen's d (Cohen 1992)), and a p-value from a permutation test:

$$\Pr_i[s(Xi, Yi, A, B) > s(X, Y, A, B)] \quad (3)$$

which shuffles the words of the target groups to determine the unlikeliness of the test statistic $s(X, Y, A, B)$, given by:

$$\sum_{x\in X} s(x, A, B) - \sum_{y\in Y} s(y, A, B) \quad (4)$$

Level 2 We introduce Level 2 of the ML-EAT, which quantifies the differential association of a single Target concept T with attributes A and B:

$$d_{A,B,T} = \frac{\text{mean}_{a\in A} u(T, a) - \text{mean}_{b\in B} u(T, b)}{\text{std\_dev}_{x\in A\cup B} u(T, x)} \quad (5)$$

where the association $u()$ for an attribute embedding $\vec{a}$ with the target group T is given by:

$$\text{mean}_{t\in T}\cos(\vec{t}, \vec{a}) \quad (6)$$

Like Level 1, Level 2 returns an effect size (Cohen's d (Cohen 1992)) and a p-value from a permutation test:

$$\Pr_i[u(Ai, Bi, T) > u(A, B, T)] \quad (7)$$

which shuffles the words of the attribute groups to determine the unlikeliness of the test statistic, defined as:

$$\sum_{a\in A} u(T, a) - \sum_{b\in B} u(T, b) \quad (8)$$

Note that Level 2 computes the differential association between a target group and two attributes, rather than an attribute and two targets. This design is intentional, as it allows the ML-EAT to answer whether a target group like Young names, for example, is associated with pleasantness or unpleasantness, without reference to another target group, such

as Old names. In this way, Level 2 generalizes the Single-Category Embedding Association Test (SC-EAT), with which Caliskan, Bryson, and Narayanan (2017) compute the differential association of a single word w (such as a job title) with two attributes (such as male female words). Consider the formula for the SC-EAT:

$$\frac{\text{mean}_{a\in A}\cos(\vec{w},\vec{a}) - \text{mean}_{b\in B}\cos(\vec{w},\vec{b})}{\text{std\_dev}_{x\in A\cup B}\cos(\vec{w},\vec{x})} \tag{9}$$

When the number of words in a target group T is equal to 1, the mean of its cosine similarities with an attribute word x is necessarily equal to the only cosine similarity computed, such that $u(T,x)$ reduces to $cos(\vec{w},\vec{x})$, and the formula for Level 2 reduces to that of the SC-EAT. This suggests that Level 2 can be used in ways analogous to the SC-EAT, and may be more robustly representative of a concept because it uses a target group T, rather than a single word w.

Level 3 Level 3 describes the distribution of cosine similarities between a target group T and an attribute A in terms of the mean $\overline{TA}$ and standard deviation σ_{TA}:

$$\overline{TA} = \frac{1}{nm}\sum_{i}^{n}\sum_{j}^{m} cos(\vec{A_i},\vec{T_j}) \tag{10}$$

$$\sigma_{TA} = \sqrt{\frac{\sum_i^n\sum_j^m (cos(\vec{A_i},\vec{T_j}) - \overline{TA})^2}{nm-1}} \tag{11}$$

Level 3 surfaces the magnitude of absolute (non-standardized) differences between groups. Including a non-standardized component in the ML-EAT provides two important insights for social scientists. First, it transparently surfaces underlying cosine similarities, which may be positive (indicating similarity between the groups), negative (indicating dissimilarity between the groups), or zero. Second, it reveals when cosine similarity may not be a meaningful measurement for an embedding space, as in the case of an anisotropic (directionally uniform) embedding (Timkey and van Schijndel 2021), wherein all vectors point in the same direction, usually due to a few especially high-magnitude dimensions (Mu and Viswanath 2018). While Level 2 permits more direct comparison between two groups of cosine similarities (*e.g.,* (X,A) and (Y,A)), it still subtracts and standardizes them, rendering unavailable any interpretation that might draw on the cosine similarities themselves.

EAT-Maps

We introduce the EAT-Map, which visualizes EAT results using a four-quadrant square. Columns correspond to the EAT target groups, with X corresponding to the first column and Y to the second. Rows correspond to the attributes, with A corresponding to the first row and B to the second. Level 2 results determine the shading of the EAT-map. For example, the upper right quadrant, associated with row A and column Y, is shaded red if $d_{A,B,Y} > 0.2$ (the minimal level defined by Cohen (1992) for a "small" effect) with a p-value > 0.05. If the X target group is differentially associated with A, the top left cell is shaded red, and the bottom left cell is shaded gray; conversely, if X is differentially associated with B, the bottom left cell is shaded red, and the top left is shaded gray. If X is associated with neither A nor B, both cells of the left column (corresponding to X) are shaded gray. This is repeated for the right column, corresponding to the Y target group. The EAT-Map intuitively visualizes a taxonomy of EAT patterns, discussed next. Figure 2 illustrates the nine possible EAT-Maps, each corresponding to a distinct EAT pattern.

EAT Patterns

We introduce EAT patterns to provide a taxonomy for describing biases quantified by Level 2. EAT patterns define EAT measurements in terms of Direction, based on whether the target groups in an EAT are individually associated (*i.e.,* exhibit a significant p-value and at least a small positive effect size) with the same attribute group, or with differing attribute groups. An EAT exhibits one of the following four categories of Direction:

- **Divergent**: X and Y are associated with differing attributes (*e.g.,* X with A, Y with B).
- **Uniform**: X and Y are associated with the same attribute (*e.g.,* X with A, and Y with A).
- **Singular**: Either X or Y is associated with an attribute, while the other target group is not (*e.g.,* X with A, Y with neither A nor B).
- **Non-Directional**: neither X nor Y is associated with either A or B.

EAT Patterns describe an EAT's target-attribute associations by prepending them to the Direction. An EAT with Singular Direction wherein the only significant Level 2 association occurs between Target Y and Attribute B exhibits a BY-Singular EAT pattern. An EAT with Uniform Direction and significant Level 2 associations both between X and A and between Y and A exhibits an A-Uniform EAT pattern. Describing an EAT that exhibits a Divergent pattern is slightly different: if significant Level 2 associations occur between X and A and between Y and B, the EAT exhibits an AB-Divergent pattern; if associations occur between X and B and between Y and A, the EAT exhibits a BA-Divergent pattern.

Societal biases quantified using Embedding Association Tests can be described more transparently by employing the vocabulary of EAT patterns; for example, describing the outcome of the Male/Female, Science/Arts EAT as AX-Singular communicates that Science (X) is differentially associated with Male (A), but Arts (Y) is associated neither with Male (A) nor with Female (B). The granular information provided by EAT patterns allows social scientists to observe more about the nature of a bias than can be interpreted via the single effect size and p-value returned by a traditional EAT. As demonstrated in the Results section and illustrated in Figure 2, nearly all EAT patterns consistent with a positively signed Level 1 effect size do in fact occur in the tests performed by Caliskan, Bryson, and Narayanan (2017), indicating that the ML-EAT can provide additional insights about diverse forms of societal bias, even where EATs have been taken previously.

Multilevel Embedding Association Test: GloVe Embeddings							
Level	Level 1	Level 2		Level 3			
EAT (Targets X/Y Attributes A/B)	A,B,X,Y	A,B,X	A,B,Y	A,X	B,X	A,Y	B,Y
Flower/Insect P/U25	1.50*	0.60*	−0.69*	.10 (.10)	.06 (.08)	.08 (.10)	.13 (.10)
Instrument/Weapon P/U25	1.53*	1.15*	−0.59*	.11 (.08)	.05 (.06)	.12 (.08)	.16 (.10)
EA/AA32 P/U25	1.40*	0.46	−0.31	.12 (.10)	.09 (.08)	-.01 (.08)	.00 (.07)
EA/AA16 P/U25	1.49*	0.35	−0.39	.11 (.10)	.08 (.08)	.00 (.08)	.01 (.08)
EA/AA16 P/U8	1.28*	1.12*	0.63	.18 (.09)	.10 (.06)	.02 (.07)	.00 (.07)
Male/Female Career/Family	1.81*	0.31	−1.39*	.18 (.09)	.16 (.05)	.09 (.09)	.23 (.06)
Math/Arts Male/Female	1.05*	0.38	−0.33	.10 (.09)	.09 (.09)	.23 (.07)	.24 (.08)
Science/Arts Male/Female	1.23*	0.83*	−0.05	.15 (.07)	.11 (.08)	.22 (.06)	.22 (.08)
Mental/Physical Temp/Perm	1.38*	−0.65	−1.20*	.24 (.12)	.29 (.12)	.18 (.10)	.32 (.15)
Young/Old P/U8	1.21*	1.09*	0.94*	.20 (.09)	.11 (.08)	.07 (.08)	.02 (.07)

Table 1: The ML-EAT reveals five EAT patterns in the tests of Caliskan, Bryson, and Narayanan (2017): AB-Divergent (Flower/Insect; Instrument/Weapon); Non-Directional (first two European/African American; Math/Arts); AX-Singular (third EA/AA; Science/Arts); BY-Singular (Career/Family; Mental/Physical); and A-Uniform (Young/Old). Level 2 shading describes significance.

Results

We apply the ML-EAT to three language technologies: static and diachronic word embeddings, GPT-2 language models and CLIP language-and-image models. Our analysis shows that a wide variety of EAT patterns occur even when Level 1 effect sizes are uniformly large, positive, and statistically significant; that Level 2 effect sizes can inform the interpretation of historical biases; and that Levels 2 and 3 of the ML-EAT can provide insight into the effects of prompting, and can surface anisotropy that may render EATs unreliable.

Empirical Analysis: GloVe Embeddings

Table 1 presents the results of the ML-EAT applied to the ten word embedding association tests of Caliskan, Bryson, and Narayanan (2017) and demonstrates that a wide variety of EAT patterns can result in a large, statistically significant Level 1 effect size (equivalent to the traditional WEAT effect size). Four tests exhibit Singular Direction:

- The EA/AA 16 P/U 8 test exhibits an **AX-Singular** pattern, indicating that European-American is significantly associated with Pleasantness, and African-American with neither Pleasantness nor Unpleasantness.
- The Science/Arts Male/Female test exhibits an **AX-Singular** pattern, indicating that Science is associated with Male, and Arts with neither Male nor Female.
- The Mental/Physical Temporary/Permanent test exhibits a **BY-Singular** pattern, indicating that Physical is associated with Permanent, and Mental with neither Temporary nor Permanent.
- The Male/Female Career/Family test exhibits a **BY-Singular** pattern, indicating that Female is associated with Family, and Male with neither Career nor Family.

The Young/Old test exhibits an **A-Uniform pattern**, indicating that Young and Old are both differentially associated with Pleasantness; however the magnitude of association is greater for Young (1.09 vs. 0.94). The most common EAT pattern observed is **Non-Directional**, exhibited by the first two European American/African American PU/25 tests and the Math/Arts Male/Female test. Only the Flowers/Insects P/U25 and Instruments/Weapons P/U25 tests exhibit an **AB-Divergent** EAT pattern, a notable finding given that discussions of EAT results often suggest this pattern (*i.e.*, X is differentially associated with A, while Y is differentially with B). That none of the results of the *social* bias tests in the GloVe embeddings exhibit an AB-Divergent EAT pattern highlights the need for descriptive reporting of EATs.

Inspection of Level 3 results reveals that small differences in cosine similarity distributions can yield large, statistically significant Level 1 effect sizes. Consider EATs exhibiting a Non-Directional pattern: in the Math/Arts, Male/Female test, the absolute difference in mean cosine similarity for Math is .01 greater with the Male attribute group (.10 vs. .09), while the absolute difference in mean cosine similarity for Arts is .01 greater with the Female attribute group (.23 vs. .24). Similarly, the mean cosine similarity for an African American names target group (Y in tests 3, 4, and 5) with any attribute group never exceeds .02 or falls below -.01, suggesting a paucity of co-occurrence data for African American names due to under-representation in the training data. This is also reflected in the non-significant Level 2 effect size A, B, Y for the African-American names target group. Nonetheless, Level 1 returns a large, significant effect size for these EATs.

That Level 1 picks up on small differences is a *benefit* of the EAT, and Level 1 is often consistent with tests of implicit bias in humans (Caliskan, Bryson, and Narayanan 2017). However, interpreting Level 1 without reference to Level 2 or 3 could lead to inaccurate conclusions about the direction of bias and the magnitude of absolute differences in underlying similarities between target and attribute groups.

Empirical Analysis: HistWords Embeddings

Among the most common uses of the EAT in computational social science is to observe change in societal biases over time (Charlesworth, Caliskan, and Banaji 2022; Borenstein et al. 2023; Betti, Abrate, and Kaltenbrunner 2023). To illustrate how the ML-EAT can inform such studies, we quantified gender bias using the Math/Arts Male/Female EAT in the HistWords embeddings (Hamilton, Leskovec,

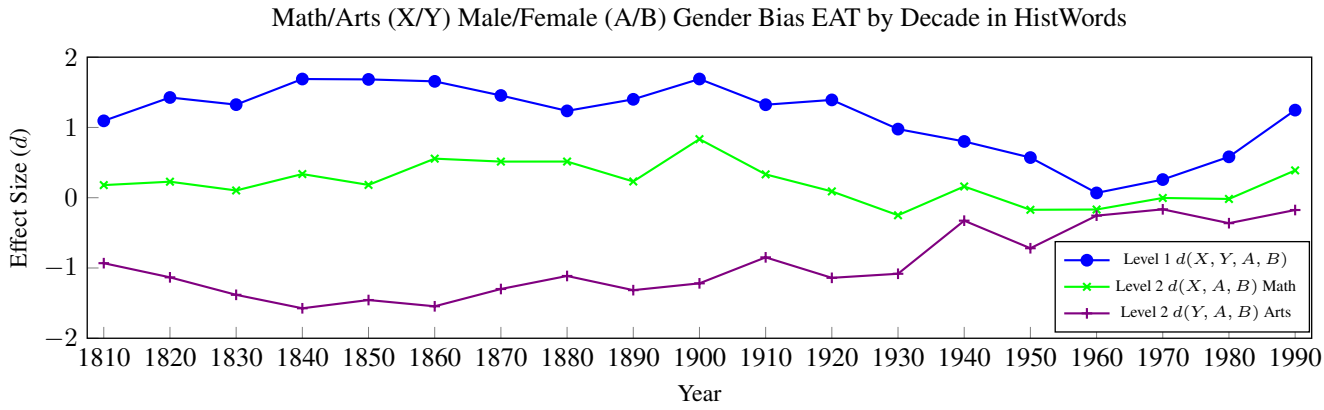

Figure 3: The ML-EAT can clarify underlying patterns of association in studies of historical bias. While Math/Arts gender bias in the 1990s appears to return to 1920s magnitudes based on Level 1, Level 2 makes clear that the underlying bias pattern (Nondirectional) has not changed - although Math does exhibit a small, non-significant effect with Male in the 1990s.

and Jurafsky 2016) from the 1810s through the 1990s (we excluded 1800 because many stimuli had zero-norm vectors, indicating insufficient co-occurrences for analysis). Figure 3 describes Level 1 effect sizes as well as Level 2 X, A, B and Y, A, B effect sizes. Note that HistWords embeddings are titled with the year that starts the decade on which they train (the 1990 embedding trains on text from 1990-1999).

By relying only on Level 1 (the traditional EAT), one might conclude that Science/Arts gender bias declined to its lowest level around 1960 ($d = .07$), only to increase sharply in the 1990s ($d = 1.25$) to a degree not observed since the 1920s. Measuring Level 2 effect sizes adds significant context: from the 1810s through the 1890s and the 1910s through the 1930s, Arts exhibits a large, statistically significant association with Female, while Math is significantly associated with neither Male nor Female - a **BY-Singular** EAT pattern. The lone outlier is the 1900s, which exhibits an **AB-Divergent** EAT pattern (Math is significantly associated with Male, Arts with Female). Starting in the 1940s, Arts is no longer significantly associated with Female, and every EAT pattern measured thereafter corresponds to **Non-Directional**. While the association of Math with Male *does increase* to $d = 0.38$ in 1990, the effect is small and not significant; Arts remains not significantly associated with either Male or Female, with $d = -0.18$. The increased observability afforded by the ML-EAT both helps to interpret changes in bias and to prevent drawing incomplete conclusions based on the Level 1 effect size.

Empirical Analysis: Prompting in CLIP

In modern zero-shot language-and-image models like CLIP, the choice of prompt can impact the cosine similarities returned by the model. For example, Radford et al. (2021) suggested adding the prefix "a photo of a **class**" to prompts when using CLIP in the zero-shot setting to improve performance when classifying images. Following the IAT, wherein human reaction times are measured in response to the appearance of pairs of individual words on a screen (Greenwald, McGhee, and Schwartz 1998), EATs typically add little context when measuring semantic associations in language technologies, and usually attempt to measure associations between individual words or images. However, some recent research using language-and-image models departs from this (Wolfe et al. 2023), employing longer prompts like that recommended by Radford et al. (2021) in order to use the model as close to the way it was intended as possible.

Level 3 of the ML-EAT can surface the impact of prompts by revealing underlying CLIP cosine similarities. We study five positively-signed, statistically significant EATs obtained from the CLIP-ViT-L14-336 model using the language stimuli of Caliskan, Bryson, and Narayanan (2017) and the image stimuli Steed and Caliskan (2021), with results in Table 2. These EATs include a Flowers-Insects P/U25 test exhibiting the **AB-Divergent** EAT pattern (*i.e.,* Flowers associated with Pleasantness, Insects with Unpleasantness); a test of White/Black P/U25 racial bias exhibiting the **B-Uniform** EAT pattern (*i.e.,* both White and Black associated with Unpleasantness); a Thin/Heavy P/U25 test of weight bias exhibiting a **BY-Singular** EAT pattern (*i.e.,* Heavy associated with Unpleasantness, Thin with neither Pleasant nor Unpleasant); a Male/Female Career/Family test of gender bias exhibiting an **A-Uniform** EAT pattern (*i.e.,* both Male and Female associated with Career); and a Science/Arts Male/Female test of gender bias exhibiting a **Non-Directional** EAT pattern (neither target associated with an attribute).

Using prompts with CLIP can affect these measurements. In accordance with the suggestion of Radford et al. (2021) and in keeping with the the IAT, we used "a picture that brings to mind [word]" as the prompt for word stimuli in both the A and B attribute groups (not all image stimuli are photographs, so we prompted with "picture" instead of "photo"). With the prompt, the mean cosine similarity increases for every target-attribute pairing. Observing the impact of prompting helps to understand the variance induced by a particular prompt, and reinforces that these EATs measure *implicit* bias: if the stimuli described exactly what was in the image, the cosine similarities would be higher, as they are proportional to probabilities in CLIP.

The CLIP measurements make clear another benefit of the

CLIP-ViT-L14-336 EAT Results - No Prompt							
Level	Level 1	Level 2		Level 3			
EAT (Targets X/Y Attributes A/B)	A,B,X,Y	A,B,X	A,B,Y	A,X	B,X	A,Y	B,Y
Flower/Insect P/U25	1.88*	0.87*	−1.03*	.15 (.02)	.14 (.02)	.13 (.02)	.15 (.02)
White/Black P/U25	1.05*	−1.27*	−1.40*	.16 (.01)	.18 (.01)	.14 (.02)	.16 (.02)
Thin/Heavy P/U25	1.58*	−0.44	−0.90*	.15 (.01)	.16 (.01)	.14 (.02)	.15 (.01)
Male/Female Career/Family	0.46*	1.66*	1.54*	.16 (.02)	.13 (.02)	.15 (.02)	.13 (.02)
Science/Arts Male/Female	0.81*	0.19	−0.33	.12 (.02)	.12 (.02)	.12 (.02)	.12 (.02)
CLIP-ViT-L14-336 EAT Results - With Prompt							
Level	Level 1	Level 2		Level 3			
EAT (Targets X/Y Attributes A/B)	A,B,X,Y	A,B,X	A,B,Y	A,X	B,X	A,Y	B,Y
Flower/Insect P/U25	1.84*	1.04*	−0.74*	.16 (.02)	.15 (.02)	.15 (.02)	.16 (.02)
White/Black P/U25	−1.20*	−0.94*	−0.87*	.17 (.01)	.18 (.01)	.15 (.01)	.16 (.01)
Thin/Heavy P/U25	1.42*	−0.24	−0.51*	.16 (.01)	.16 (.01)	.16 (.01)	.16 (.01)
Male/Female Career/Family	0.29	1.25*	1.09*	.17 (.02)	.15 (.02)	.16 (.02)	.15 (.02)
Science/Arts Male/Female	0.49	−0.53	−0.80	.13 (.02)	.14 (.02)	.14 (.02)	.15 (.02)

Table 2: ML-EAT results on CLIP-ViT-L14-336 demonstrate that changing the prompt can change the *sign* of the Level 1 effect size, while EAT patterns (defined by Level 2) remain unchanged. Note target groups are images and attribute groups are text.

ML-EAT. Of the five Level 1 measurements, two are significant only in the absence of the prompt, and the White/Black P/U25 test changes the direction of association. Without the prompt, Level 1 indicates a large, significant valence bias favoring White over Black; with the prompt, Level 1 indicates a large, significant valence bias favoring Black over White. However, the EAT patterns for all five tests, which are based on Level 2 effect sizes, remain the same regardless of whether the prompt is present. This may not always be the case, but it illustrates the importance of being able to observe the measurements that underlie the top level EAT d-score.

Empirical Analysis: Anisotropy in GPT-2

Table 3 describes results of the ML-EAT in both the GPT-2 Base model (124 million parameters) and the GPT-2 XL model (1.5 billion parameters) measured using the prompting approach of May et al. (2019), wherein the model receives the prompt "This is [stimulus]" to accord with its training objective. We focus not on the EAT patterns in this case, but on the Level 3 results, which provide evidence of anisotropy (directional uniformity, based on cosine similarity close to 1.0). Given that prior work finds that high anisotropy obscures the semantic properties of contextual word embeddings (Mu and Viswanath 2018; Timkey and van Schijndel 2021), one might avoid relying on the Level 1 and Level 2 measurements for which these cosine similarities are components. Cosine similarities in GPT-2 XL, on the other hand, may exhibit mild anisotropy, but they also exhibit more variance than the smaller GPT-2 model. While EAT patterns are mostly nondirectional, Level 1 effects are large and significant in the XL model, consistent with results from static word embeddings and with the studies of implicit bias in human subjects.

Discussion

Reporting outcomes using the ML-EAT can increase the transparency of research that employs EATs. Rather than attempting to explain the meaning of a single summary statistic and significance test, researchers can drawn on a categorical and visual vocabulary with which to communicate the findings of their work. Given the variance observed in the Level 2 results obtained using stimuli employed in prior studies (even with uniformly large Level 1 effect sizes), describing intermediate results via the ML-EAT might be adopted as a best practice when reporting the outcome of an EAT.

Motivating Interpretable Bias Measurement

Studies employing the EAT inform how social scientists understand human attitudes (Charlesworth, Caliskan, and Banaji 2022), and ongoing work uses these measurements to *predict* human bias (Bhatia and Walasek 2023). Ensuring that findings related to societal or domain-specific (*e.g.,* legal, medical, etc.) bias are well-understood is essential not only for the integrity of the scientific record but for the decisions societies may make on the basis of that data. Scholars including Greenwald et al. (2022) have suggested that implicit bias might be approached as a public health problem, and mitigated using preventative approaches "to disable the path from implicit biases to discriminatory outcomes." Where an EAT is presented as evidence of an implicit bias in need of redress through public health measures, researchers would be well-served by transparent and interpretable methods.

Where the ML-EAT informs how intrinsic bias is interpreted, it may also provide information for how bias might be addressed in the embedding itself. In the case of the tests of racial bias (EA/AA Names) in the GloVe embeddings, Level 2 indicates that there is no significant association of African American names with pleasantness or unpleasantness, and Level 3 indicates that this results from a lack of similarity (due to lack of co-occurrence) with words in either attribute A or B. This suggests that bias quantified by the EAT might be mitigated via more diverse and representative training data, rather than by aggressively pruning the dataset, a process which can exclude diverse voices (Dodge et al. 2021).

Improving the Robustness of the EAT

The ML-EAT also helps to improve the robustness of EAT-based measurements by introducing a generalization of the

GPT-2 Base Full EAT Results							
Level	Level 1	Level 2		Level 3			
EAT (Targets X/Y Attributes A/B)	A,B,X,Y	A,B,X	A,B,Y	A,X	B,X	A,Y	B,Y
Flower/Insect P/U25	0.42	−0.30	−0.54*	.99 (.01)	.99 (.01)	.99 (.00)	.99 (.00)
Instrument/Weapon P/U25	−0.20	−0.16	−0.01	.99 (.00)	.99 (.00)	.99 (.01)	.99 (.01)
EA/AA32 P/U25	0.23	0.27	0.30	.97 (.01)	.97 (.01)	.98 (.01)	.98 (.01)
EA/AA16 P/U25	−0.19	0.18	0.37	.97 (.01)	.97 (.01)	.98 (.01)	.98 (.01)
EA/AA16 P/U8	0.04	−0.22	−0.34	.97 (.01)	.97 (.01)	.98 (.01)	.98 (.01)
Male/Female Career/Family	0.08	1.31*	1.30*	.98 (.01)	.97 (.01)	.98 (.01)	.97 (.01)
Math/Arts Male/Female	−0.17	−0.58	−0.46	.99 (.01)	.99 (.01)	.99 (.01)	.99 (.00)
Science/Arts Male/Female	−0.44	−0.24	−0.18	.99 (.01)	.99 (.01)	.99 (.01)	.99 (.00)
Mental/Physical Temporary/Permanent	−1.20*	0.81	0.67	.99 (.00)	.99 (.01)	.99 (.00)	.98 (.01)
Young/Old P/U8	−0.44	−0.31	−0.20	.97 (.01)	.97 (.01)	.98 (.01)	.98 (.01)
GPT-2 XL Full EAT Results							
Level	Level 1	Level 2		Level 3			
EAT (Targets X/Y Attributes A/B)	A,B,X,Y	A,B,X	A,B,Y	A,X	B,X	A,Y	B,Y
Flower/Insect P/U25	1.65*	0.04	−0.69*	.29 (.05)	.29 (.06)	.28 (.06)	.31 (.07)
Instrument/Weapon P/U25	0.95*	−0.02	−0.28	.28 (.05)	.28 (.05)	.31 (.06)	.32 (.07)
EA/AA32 P/U25	0.91*	0.37	0.24	.23 (.06)	.21 (.05)	.21 (.06)	.20 (.06)
EA/AA16 P/U25	0.64*	0.28	0.18	.22 (.05)	.21 (.05)	.22 (.07)	.21 (.06)
EA/AA16 P/U8	0.65*	0.14	−0.11	.25 (.04)	.25 (.03)	.24 (.06)	.25 (.06)
Male/Female Work/Home	1.17*	−0.81	−1.02*	.18 (.04)	.22 (.06)	.18 (.05)	.23 (.06)
Math/Arts Male/Female	0.19	−0.57	−0.56	.31 (.06)	.34 (.06)	.33 (.07)	.36 (.07)
Science/Arts Male/Female	0.33	−0.36	−0.40	.31 (.06)	.33 (.06)	.31 (.07)	.34 (.07)
Mental/Physical Temporary/Permanent	1.52*	0.75	0.46	.49 (.09)	.41 (.10)	.34 (.07)	.32 (.05)
Young/Old P/U8	1.27*	0.56	0.00	.27 (.04)	.25 (.03)	.24 (.05)	.24 (.04)

Table 3: Level 3 of the ML-EAT surfaces the directional uniformity of contextualized embeddings in GPT-2 base, which results in low variance in cosine similarity and inconsistent Level 1 measurements. However, Level 1 measurements are consistent with societal biases in the XL model, which has variance in cosine similarities comparable to that observed in static word embeddings.

SC-EAT in Level 2. Generalizing this test such that the target group can be defined with additional words improves the robustness of single-target tests, which are otherwise dependent on the conceptual representativeness of a single word. Moreover, the ML-EAT is modular enough to use definitions of association other than cosine similarity. For example, one could adopt algebraic definition, following Bolukbasi et al. (2016), but still use the framework of the ML-EAT to provide transparency at multiple levels of measurement.

Intrinsic Bias and Application Bias

Prior work documenting limitations of EATs has largely focused on evidence that intrinsic biases do not transfer to downstream tasks (Goldfarb-Tarrant et al. 2020). While it is not our central concern, we note that, in cases where cosine similarity has an explicitly defined function in a model, biases measured transparently using well-designed EATs will necessarily transfer downstream. This occurs when models like CLIP are used in a zero-shot setting, such that a cosine similarity between text and image is converted into a *probability* for use in classification (Radford et al. 2021). While intrinsic measurements may not predict application bias in many cases, there remain NLP applications that motivate the transparency and interpretability of EAT measurements.

Limitations and Future Work

While the ML-EAT renders bias more observable, careful curation of stimuli remains necessary to ensure validity (Caliskan et al. 2022). Ensuring that those stimuli are globally representative is an ongoing challenge for studies employing EATs, as recent research contends that stimuli used in most EATs reflect a western-centric bias that fails to capture biases faced by indigenous populations around the world (Yogarajan, Dobbie, and Gouk 2023). Moreover, while the ML-EAT is modular with differing mathematical definitions of bias, its levels are not adaptable for some modified versions of the EAT, such as the CEAT, which computes associations using thousands of sentences (Guo and Caliskan 2021). Future work might extend the ML-EAT to such tests and uncover new patterns of bias. Finally, our research addresses limitations of observability of bias in EATs, rather than downstream propagation of intrinsic bias.

Conclusion

We introduced the ML-EAT, a three-level measurement of intrinsic bias intended to improve the transparency and observability of bias measurement in social science, and appropriate for technologies ranging from static and diachronic word embeddings to zero-shot language-and-image models. We further introduced a taxonomy of nine distinct EAT patterns which we showed occur in prior EATs applied to word embeddings, alongside the EAT-Map, an intuitive visualization for EAT patterns. The ML-EAT provides greater transparency when employing language technologies to understand human minds and societies, an increasing concern as such measurements are used not only to observe human bias, but to predict it (Bhatia and Walasek 2023), and perhaps even to try to prevent it (Greenwald et al. 2022).

Researcher Positionality

Two of the authors of this research have an extensive background in machine learning and quantitative studies of AI bias, and a third author has extensive experience with human-computer interaction, including statistical studies of deceptive design. We sought to include a variety of perspectives on this project, as the method we intended to develop needed to be both statistically rigorous, yet also approachable and interpretable for researchers hoping to study bias in AI.

Ethical Considerations

We note that, while we believe reporting results using the ML-EAT will help to make bias research more transparent and interpretable, simply using the ML-EAT rather than the WEAT will not guarantee full transparency in research practices. Researchers must also choose ethical ways of selecting stimuli for WEAT tests, and for determining the number of stimuli to include, which can impact the statistical power of the test. Though pre-registration is sometimes employed for psychological experiments, including those involving word embeddings, the easy availability of language models may render this approach less effective than it is with human subjects experiments, which carry much more notable startup costs. Future work might consider ethical approaches to social scientific experiment design with modern language technologies.

Adverse Impacts

While we have not produced any new technology in this work, individuals could use our method for ends we have not intended, such as exploiting biases identified with the ML-EAT to further marketing campaigns or to produce misinformation targeted to societal vulnerabilities. We hope and expect that most uses of the method will be to support more transparent and interpretable studies of bias in AI.

References

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.

Bai, X.; Wang, A.; Sucholutsky, I.; and Griffiths, T. L. 2024. Measuring Implicit Bias in Explicitly Unbiased Large Language Models. *arXiv preprint arXiv:2402.04105*.

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

Betti, L.; Abrate, C.; and Kaltenbrunner, A. 2023. Large scale analysis of gender bias and sexism in song lyrics. *EPJ Data Science*, 12(1): 10.

Bhatia, S.; and Walasek, L. 2023. Predicting implicit attitudes with natural language data. *Proceedings of the National Academy of Sciences*, 120(25): e2220726120.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29: 4349–4357.

Borenstein, N.; Stańczak, K.; Rolskov, T.; Perez, N. d. S.; Käfer, N. K.; and Augenstein, I. 2023. Measuring Intersectional Biases in Historical Documents. *arXiv preprint arXiv:2305.12376*.

Burnell, R.; Schellaert, W.; Burden, J.; Ullman, T. D.; Martinez-Plumed, F.; Tenenbaum, J. B.; Rutar, D.; Cheke, L. G.; Sohl-Dickstein, J.; Mitchell, M.; et al. 2023. Rethink reporting of evaluation results in AI. *Science*, 380(6641): 136–138.

Cabello, L.; Jørgensen, A. K.; and Søgaard, A. 2023. On the Independence of Association Bias and Empirical Fairness in Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 370–378.

Caliskan, A.; Ajay, P. P.; Charlesworth, T.; Wolfe, R.; and Banaji, M. R. 2022. Gender bias in word embeddings: a comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 156–170.

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

Charlesworth, T. E.; Caliskan, A.; and Banaji, M. R. 2022. Historical representations of social groups across 200 years of word embeddings from Google Books. *Proceedings of the National Academy of Sciences*, 119(28): e2121798119.

Charlesworth, T. E.; Yang, V.; Mann, T. C.; Kurdi, B.; and 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.

Chen, M.; Radford, A.; Child, R.; Wu, J.; Jun, H.; Luan, D.; and Sutskever, I. 2020a. Generative pretraining from pixels. In *International conference on machine learning*, 1691–1703. PMLR.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020b. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Cobert, J.; Mills, H.; Lee, A.; Gologorskaya, O.; Espejo, E.; Jeon, S. Y.; Boscardin, J.; Heintz, T. A.; Kennedy, C.; Ashana, D.; et al. 2024. Measuring Implicit Bias in ICU Notes Using Word-Embedding Neural Network Models. *Chest*.

Cohen, J. 1992. Statistical power analysis. *Current directions in psychological science*, 1(3): 98–101.

Dodge, J.; Sap, M.; Marasovic, A.; Agnew, W.; Ilharco, G.; Groeneveld, D.; and Gardner, M. 2021. Documenting the english colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*.

Durrheim, K.; Schuld, M.; Mafunda, M.; and Mazibuko, S. 2023. Using word embeddings to investigate cultural biases. *British Journal of Social Psychology*, 62(1): 617–629.

Dutta, S.; Srivastava, P.; Solunke, V.; Nath, S.; and KhudaBukhsh, A. R. 2023. Disentangling societal inequality from model biases: gender inequality in divorce court proceedings. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 5959–5967.

Ethayarajh, K.; Duvenaud, D.; and Hirst, G. 2019. Understanding undesirable word embedding associations. *arXiv preprint arXiv:1908.06361*.

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

Goldfarb-Tarrant, S.; Marchant, R.; Sánchez, R. M.; Pandya, M.; and Lopez, A. 2020. Intrinsic bias metrics do not correlate with application bias. *arXiv preprint arXiv:2012.15859*.

Greenwald, A. G.; Dasgupta, N.; Dovidio, J. F.; Kang, J.; Moss-Racusin, C. A.; and Teachman, B. A. 2022. Implicit-bias remedies: Treating discriminatory bias as a public-health problem. *Psychological Science in the Public Interest*, 23(1): 7–40.

Greenwald, A. G.; McGhee, D. E.; and Schwartz, J. L. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6): 1464.

Guan, L.; Shi, W.; Li, Q.; Oktavianus, J.; and Wu, M. 2024. Have color representations in books changed over the past 200 years? An empirical analysis based on the Google Books Ngram corpus. *Color Research & Application*, 49(1): 65–78.

Guo, W.; and Caliskan, A. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 122–133.

Hamilton, W. L.; Leskovec, J.; and Jurafsky, D. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1489–1501.

Hausladen, C. I.; Knott, M.; Perona, P.; and Camerer, C. 2023. Causal analysis of social bias in CLIP.

Kennedy, B.; Ashokkumar, A.; Boyd, R. L.; and Dehghani, M. 2021. Text analysis for psychology: Methods, principles, and practices.

Kurita, K.; Vyas, N.; Pareek, A.; Black, A. W.; and Tsvetkov, Y. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Leach, S.; Kitchin, A. P.; and Sutton, R. M. 2023. Word embeddings reveal growing moral concern for people, animals and the environment. *British Journal of Social Psychology*, 62(4): 1925–1938.

Lewis, M.; and Lupyan, G. 2020. Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature human behaviour*, 4(10): 1021–1028.

Lin, Y.; Michel, J.-B.; Lieberman, E. A.; Orwant, J.; Brockman, W.; and Petrov, S. 2012. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, 169–174.

Manzini, T.; Lim, Y. C.; Tsvetkov, Y.; and Black, A. W. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*.

Matthews, S.; Hudzina, J.; and Sepehr, D. 2022. Gender and Racial Stereotype Detection in Legal Opinion Word Embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 12026–12033.

May, C.; Wang, A.; Bordia, S.; Bowman, S.; and Rudinger, R. 2019. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 622–628.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Morehouse, K.; Rouduri, V.; Cunningham, W.; and Charlesworth, T. 2023. Traces of Human Attitudes in Contemporary and Historical Word Embeddings (1800-2000). *Research Square Preprint*.

Mu, J.; and Viswanath, P. 2018. All-but-the-Top: Simple and Effective Postprocessing for Word Representations. In *International Conference on Learning Representations*.

Mukherjee, A.; Raj, C.; Zhu, Z.; and Anastasopoulos, A. 2023. Global Voices, Local Biases: Socio-Cultural Prejudices across Languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 15828–15845.

Napp, C. 2023. Gender stereotypes embedded in natural language are stronger in more economically developed and individualistic countries. *PNAS nexus*, 2(11): pgad355.

Omrani Sabbaghi, S.; Wolfe, R.; and Caliskan, A. 2023. Evaluating biased attitude associations of language models in an intersectional context. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 542–553.

Orgad, H.; Goldfarb-Tarrant, S.; and Belinkov, Y. 2022. How gender debiasing affects internal model representations, and why it matters. *arXiv preprint arXiv:2204.06827*.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22.

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

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, 28492–28518. PMLR.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Rios, A.; Joshi, R.; and Shin, H. 2020. Quantifying 60 years of gender bias in biomedical research with word embeddings. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, 1–13.

Ross, C.; Katz, B.; and Barbu, A. 2020. Measuring social biases in grounded vision and language embeddings. *arXiv preprint arXiv:2002.08911*.

Schmahl, K. G.; Viering, T. J.; Makrodimitris, S.; Naseri Jahfari, A.; Tax, D.; and Loog, M. 2020. Is Wikipedia succeeding in reducing gender bias? Assessing changes in gender bias in Wikipedia using word embeddings. In Bamman, D.; Hovy, D.; Jurgens, D.; O'Connor, B.; and Volkova, S., eds., *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, 94–103. Online: Association for Computational Linguistics.

Shanahan, M.; McDonell, K.; and Reynolds, L. 2023. Role play with large language models. *Nature*, 623(7987): 493–498.

Slaughter, I.; Greenberg, C.; Schwartz, R.; and Caliskan, A. 2023. Pre-trained Speech Processing Models Contain Human-Like Biases that Propagate to Speech Emotion Recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 8967–8989.

Steed, R.; and Caliskan, A. 2021. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 701–713.

Sunsay, C. 2023. A historical evaluation of the disease avoidance theory of xenophobia. *Plos one*, 18(12): e0294816.

Timkey, W.; and van Schijndel, M. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. *arXiv preprint arXiv:2109.04404*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

Wolfe, R.; Banaji, M. R.; and Caliskan, A. 2022. Evidence for Hypodescent in Visual Semantic AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

Wolfe, R.; and Caliskan, A. 2022a. American==White in Multimodal Language-and-Image AI. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*.

Wolfe, R.; and Caliskan, A. 2022b. VAST: The Valence-Assessing Semantics Test for Contextualizing Language Models. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*.

Wolfe, R.; Yang, Y.; Howe, B.; and Caliskan, A. 2023. Contrastive language-vision ai models pretrained on web-scraped multimodal data exhibit sexual objectification bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1174–1185.

Xu, Y.; Wang, S.; Li, P.; Luo, F.; Wang, X.; Liu, W.; and Liu, Y. 2023. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*.

Yogarajan, V.; Dobbie, G.; and Gouk, H. 2023. Effectiveness of Debiasing Techniques: An Indigenous Qualitative Analysis.