Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)
Early Career Track

Artificial Intelligence, Bias, and Ethics

Aylin Caliskan
University of Washington
aylin@uw.edu

Abstract

Although ChatGPT attempts to mitigate bias, when instructed to translate the gender-neutral Turkish sentences "O bir doktor. O bir hemşire" to English, the outcome is biased: "He is a doctor. She is a nurse." In 2016, we have demonstrated that language representations trained via unsupervised learning automatically embed implicit biases documented in social cognition through the statistical regularities in language corpora. Evaluating embedding associations in language, vision, and multi-modal vision-language models reveals that large-scale sociocultural data is a source of implicit human biases regarding gender, race or ethnicity, skin color, ability, age, sexuality, religion, social class, and intersectional associations. The study of gender bias in language, vision, vision-language, and generative AI has highlighted the sexualization of women and girls in AI, while easily accessible generative AI models such as text-to-image generators amplify bias at scale. As AI increasingly automates tasks that determine life's outcomes and opportunities, the ethics of AI bias has significant implications for human cognition, society, justice, and the future of AI. Thus, it is necessary to advance our understanding of the depth, prevalence, and complexities of bias in AI to mitigate it both in machines and society.

1 Introduction

Large-scale language corpora of human-produced texts are sources of implicit associations and biases [Caliskan *et al.*, 2016a; Caliskan *et al.*, 2017]. Consequently, natural language processing models that train language representations via unsupervised learning discover indirect patterns in data and learn implicit associations. Similarly, contextualized representations of artificial intelligence (AI) models in the language, vision, and vision-language modalities trained on large-scale sociocultural data embed implicit associations and biases [Guo and Caliskan, 2021; Steed and Caliskan, 2021a; Wolfe and Caliskan, 2022e; Wolfe and Caliskan, 2022a]. These associations propagate to AI applications and manifest in the outputs of generative AI models. Extending beyond the scope of computer science, information science, and robotics, these advancements hold significance within social scientific disciplines, including, psychology, cognitive science, linguistics, political science, and the humanities. The scientific fact that unsupervised AI models, trained on sociocultural data, inherently acquire and exhibit implicit associations and biases holds profound implications for humans, society, justice, policy, and the future of AI [Pandey and Caliskan, 2021; Caliskan, 2021; Caliskan and Steed, 2022].

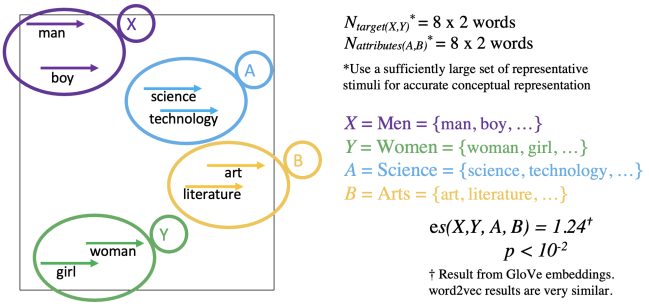

Figure 1: WEAT quantifying the effect size (*es* in d) of bias: standardized differential association of two sets of social groups (*men* and *women*) with two concepts (*science* and *arts*). $d = 1.24$ is a large positive effect size relatively associating men with science and women with arts. On the left side of this figure, sets of 300 dimensional word embeddings from the geometric space of language are approximately projected to two dimensions for illustration purposes.

As AI models become part of humans' daily lives by automating consequential decision making tasks, presenting easily accessible human-AI interaction settings, and generating synthetic data with a lasting presence, developing bias evaluation methods and enhancing AI transparency becomes a long term research endeavor in trustworthy AI that advances the science of bias in AI and society. This pursuit also represents the necessary initial step toward identifying and mitigating bias in both AI systems and broader societal contexts.

The geometric space of language contains low-dimensional word embeddings for machines to effectively and efficiently process language. These embeddings are trained on word co-occurrence statistics derived from large-scale language corpora to represent language, information, semantics, and syntax. Developing the Word Embedding Association Test (WEAT), illustrated in Figure 1, revealed that word embed-

Test	**Effect Size** (es measured in d)	***p*-value**
WEAT	$es(X, Y, A, B) = \frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std-dev}_{w \in X \cup Y} s(w, A, B)}$	$\Pr_i[s(X_i, Y_i, A, B) > s(X, Y, A, B)]$
SC-WEAT	$es(\vec{w}, A, B) = \frac{\text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})}{\text{std-dev}_{x \in A \cup B} \cos(\vec{w}, \vec{x})}$	$\Pr_i[s(\vec{w}, A_i, B_i) > s(\vec{w}, A, B)]$

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

Cosine similarity: $\cos(\vec{a}, \vec{b})$ denotes the cosine of the angle between the vectors $\vec{a}$ and $\vec{b}$.

Target words: $X = [\vec{x_1}, \vec{x_2}, \ldots, \vec{x_m}]$ and $Y = [\vec{y_1}, \vec{y_2}, \ldots, \vec{y_m}]$ are the two equal-sized (m), sufficiently large, and representative sets of target stimuli. Target words represent concepts, such as social groups.

Attribute words: $A = [\vec{a_1}, \vec{a_2}, \ldots, \vec{a_n}]$ and $B = [\vec{b_1}, \vec{b_2}, \ldots, \vec{b_n}]$ are the two equal-sized (n), sufficiently large, and representative sets of attributes. Attribute words represent concepts, such as valenced pleasant and unpleasant terms.

SC-WEAT: $\vec{w}$ is the single target stimulus in SC-WEAT.

Permutation test: $(X_i, Y_i)_i$ and $(A_i, B_i)_i$ denote all the partitions of $X \cup Y$ and $A \cup B$ into two sets of equal size. Random permutations of these sets represent the null hypothesis as if the biased associations did not exist so that we can perform a statistical significance test by measuring the unlikelihood of the null hypothesis, given the effect size of WEAT or SC-WEAT.

Effect size es in d: Cohen's d is a commonly used measure of effect size, with values of 0.20, 0.50, and 0.80 representing small, medium, and large effect sizes, respectively. The sign of the effect size indicates the direction of association.

Ongoing work: As WEAT and SC-WEAT have been generalizing to other modalities and modeling architectures via association tests, the methods are being improved by (1) overcoming limitations inherent in data for inclusive representation of social groups and concepts, (2) adapting concept representation and extraction approaches to modalities and architectures, (3) enhancing the representation of concepts across modalities for accuracy and robustness, and (4) taking into account the effects of contextualization, anisotropy, polysemy, negation, frequency, limitations of effect size measurement approaches, and information overlap quantification metrics such as cosine similarity.

Table 1: WEAT and SC-WEAT effect size equations and their corresponding statistical significance in p-values.

dings capture human-like implicit associations and biases by leveraging the statistical regularities in large-scale language corpora [Caliskan *et al.*, 2016a; Caliskan *et al.*, 2017].

The study of implicit associations concerning social groups and concepts have been extensively explored in the domains of cognition and psychology over several decades. To assess implicit biases present in human cognition, researchers have utilized the Implicit Association Test (IAT) [Greenwald and Banaji, 1995; Greenwald *et al.*, 1998]. This computerized task involves pairing words or pictures to measure standardized differential reaction latencies when categorizing stimuli in settings that are either stereotype-congruent or stereotype-incongruent. For instance, the IAT measures whether individuals are faster in associating the concepts of men and science or women and science, compared to reaction latencies in associating men and arts vs. women and arts. Drawing inspiration from the measurement of associations in cognitive psychology, we developed the WEAT outlined in Table 1. The WEAT utilizes cosine distance between word embeddings (as an analog to reaction latency in humans) and adapts the standardized differential concept association method to quantify the magnitude of bias in machines with an effect size (d). As an example, when given two sets of words representing social groups (e.g., men and women) and two sets of attribute words representing concepts (e.g., science and arts), Figure 1 visually depicts which social group exhibits a stronger association with the representation of the concept of science. Additionally, the Single-Category WEAT (SC-WEAT), which forms a fundamental component of WEAT, enables the quantification of the relative association between a single stimulus and two concepts [Caliskan *et al.*, 2017; Toney-Wails and Caliskan, 2021].

The development of WEAT has served as a pivotal tool in empirically studying bias and ethics in AI. This research has transformed our understanding of bias and cognition by demonstrating that large-scale sociocultural data is a source of associations and biases. To extend the investigation beyond language to other modalities and architectures, including state-of-the-art large language models, vision models, and vision-language models, we have introduced embedding association tests (EATs) [Guo and Caliskan, 2021; Steed and Caliskan, 2021a; Wolfe and Caliskan, 2022a; Wolfe and Caliskan, 2022e]. The collective findings from these modalities indicate that AI models trained on large-scale sociocultural data acquire implicit biases pertaining to gender, race or ethnicity, skin color, social class, ability, age, body weight, sexuality, religion, and intersectional associations [Omrani Sabbaghi and Caliskan, 2022; Omrani Sabbaghi *et al.*, 2023; Ghosh and Caliskan, 2023]. EATs not only replicate associations and biases that have been extensively documented in social psychology and implicit cognition research over the past few decades but also provide an evaluation method to study the evolution of associations and biases, the magnitude and characteristics of bias in AI models trained on large-scale data, and bias mitigation strategies.

A significant concern lies in the strong association of biases in AI with gender identity signals pertaining to women, non-binary individuals, fluid representations of gender, and men. Transparency enhancing approaches employed in language, vision, and vision-language models have revealed a prevailing association of women with the concepts of appearance, sexualized content, cooking, and the kitchen [Steed and

Caliskan, 2021a; Caliskan *et al.*, 2022; Wolfe *et al.*, 2023]. Furthermore, gender biases emerge when ChatGPT is instructed to translate from gender-neutral languages to grammatically gendered languages, and ChatGPT ignores gender-neutral pronouns of individuals in translation tasks [Ghosh and Caliskan, 2023], similar to the findings we have demonstrated in machine translation systems in 2016 [Caliskan *et al.*, 2016b]. The use of generative AI models such as Stable Diffusion in text-to-image generation leads to the creation of sexualized images of girls and women, raising potential legal and ethical concerns. Furthermore, models such as Stable Diffusion and Dall-E perpetuate heteronormative societal defaults, while amplifying complex biases with far-reaching global implications that are challenging to address [Bianchi *et al.*, 2023]. Considering the extensive utilization of these easily-accessible large-scale generative AI models, it is imperative to prioritize the evaluation of biases, the associated harms in representation, allocation, and safety, as well as the development of context-aligned sociotechnical strategies for effectively handling bias. This prioritization is necessary for the advancement of trustworthy AI.

2 Evaluating Associations and Biases in AI

We have developed bias evaluation methods and transparency enhancing approaches to effectively identify, quantify, and characterize associations, biases, and real-world statistics acquired by language, vision, and vision-language models [Guo and Caliskan, 2021; Steed and Caliskan, 2021a; Steed and Caliskan, 2021b; Toney-Wails and Caliskan, 2021; Wolfe and Caliskan, 2021; Wolfe and Caliskan, 2022a; Wolfe and Caliskan, 2022b; Wolfe and Caliskan, 2022c; Wolfe and Caliskan, 2022d; Wolfe and Caliskan, 2022e; Wolfe *et al.*, 2022; Caliskan *et al.*, 2022; Mei *et al.*, 2023; Omrani Sabbaghi *et al.*, 2023]. EATs replicated implicit associations and biases documented in psychology. These findings suggest that machines embed the fundamental aspects of cognition as they automatically learn human patterns of association from large-scale sociocultural data without explicit supervision. The WEAT also validated the IAT that has been taken by millions of individuals globally over the past two decades. By systematically examining associations at scale across various modalities, we have gained valuable insights into how visual and linguistic concepts represented in datasets collected from the internet are skewed when compared to real-world societal representations and statistics. The findings revealed bias amplification in AI and prevalent association of certain potentially harmful or disadvantaging concepts with specific social groups, further deepening our understanding of the complexities surrounding biases in AI systems.

One of our contributions is the introduction of the Single-Category WEAT (SC-WEAT), which enables the measurement of relative associations of a single stimulus with sets of attributes, providing nuanced explanations of implicit associations. These methods allow for principled large-scale studies to analyze how the linguistic and visual information processed by machines shape their concepts, associations, and biases, and ultimately impact society. Additionally, our research has extended to exploring the sources and evolution of associations, norms, social group biases, and notions of equity in both natural language processing and society as a whole [Toney *et al.*, 2021; Charlesworth *et al.*, 2022].

We have made contributions to the understanding of biases in word embeddings and their relation to human cognition [Caliskan and Lewis, 2021]. Furthermore, our research on AI, bias, and ethics has had an impact on public policy and AI regulation, with citations in policy documents around the world [Caliskan, 2021; Caliskan and Steed, 2022].

We introduced a novel intrinsic evaluation task called ValNorm, which assesses the alignment between static word embedding associations and human valence norms. Our specific focus was on valence, a principal dimension of affect of affect closely linked to attitudes. This evaluation involves quantifying the affective valence dimension of words, which have been rated by individuals [Toney-Wails and Caliskan, 2021]. Through ValNorm, we measured the valence of non-discriminatory and non-social group word sets across corpora spanning over two centuries and seven different languages, thereby uncovering widely-shared valence norm associations. Conversely, our examination of gender stereotypes has revealed varying degrees of social biases across different languages. Furthermore, we have underscored the importance of accounting for grammatical gender signals when evaluating gender bias in models trained on diverse language structures [Omrani Sabbaghi and Caliskan, 2022].

3 Generalization Across Modalities and Architectures

We have adapted the EATs across different modalities and architectures, including unsupervised and generative language, vision, and vision-language models. The results obtained from bias evaluation methods consistently demonstrate the existence of significant and complex biases within AI systems. As AI models continue to improve in terms of their accuracy and performance across various evaluation metrics, they also exhibit an increased ability to capture nuanced and complex associations and biases that are inherently embedded within the structure of large-scale training datasets [Toney-Wails and Caliskan, 2021].

3.1 Language Models

Building on ValNorm, we introduced the Valence-Assessing Semantics Test (VAST) as an intrinsic evaluation task for contextualized word embeddings of language models to assess their semantic quality and unique properties in relation to bias measurements, languages, contextualization, tokenization, downstream tasks, and language model-specific geometry [Wolfe and Caliskan, 2022e].

To measure biases in language models more comprehensively, we developed the Contextualized Embedding Association Test (CEAT), which incorporates a random-effects model to quantify overall bias magnitude without relying on sentence templates [Guo and Caliskan, 2021]. CEAT successfully identified social and intersectional biases, revealing biased representations in all the English language models studied. Additionally, we introduced Intersectional Bias Detection (IBD) and Emergent Intersectional Bias Detection

(EIBD) methods to automatically identify and measure intersectional and emergent biases in both static and contextualized word embeddings. Our findings highlighted the high magnitudes of biases at the intersection of race and gender for intersectional identities, such as African American females and Mexican American females, across language models.

3.2 Computer Vision Models

The field of computer vision has made significant progress in leveraging vast datasets of images obtained from the internet, enabling the development of general-purpose image representations for various tasks, ranging from image generation to face recognition. To investigate whether unsupervised computer vision models, such as iGPT, inherently acquire implicit associations and incorporate social biases that may have detrimental effects on representation, generation, and downstream applications, we have extended our bias measurement methods, originally developed for language representations, to the computer vision domain.

Our findings indicate that state-of-the-art unsupervised computer vision models trained on ImageNet, a widely used benchmark dataset consisting of internet-sourced images, automatically acquire biases associated with race or ethnicity, gender, skin-tone, weight, religion, and intersectionality, as these attributes are commonly portrayed in stereotypical ways across the web [Steed and Caliskan, 2021a]. By leveraging iEAT, we have replicated human associations and biases documented by the IAT, spanning from seemingly innocuous widely shared associations such as flower-pleasant/insect-unpleasant to potentially harmful biases related to race and gender. Additionally, the results closely align with three hypotheses from social psychology pertaining to intersectional bias encompassing race and gender. Furthermore, when conducting image generation experiments with synthetic image segments, we observed that women are predominantly associated with sexualized depictions.

3.3 Vision-Language Models

We compared the geometry and semantic properties of contextualized English language representations formed by GPT-2 and CLIP (Contrastive Language Image Pretraining), a zero-shot multimodal image classifier that leverages the GPT-2 architecture to encode image captions in a multi-modal vision-language space [Wolfe and Caliskan, 2022b]. In these types of vision-language models, the language modality implicitly supervises the training of the visual modality. Our findings revealed that contrastive visual semantic pretraining mitigates the anisotropy observed in contextualized word embeddings from GPT-2. This indicates that visual semantic pretraining not only improves the organization of visual representations, as measured by ValNorm, but also enhances the encoding of semantically meaningful representations in language, at both the word and sentence levels.

We evaluated CLIP, SLIP, and BLIP, state-of-the-art vision-language AI models, for a bias observed in psychology: equating American identity with Whiteness. EATs using standardized images of self-identified Asian, Black, Latina/o, and White individuals from the Chicago Face Database revealed stronger associations of White individuals with collective in-group words compared to Asian, Black, or Latina/o individuals. Assessments of American identity aspects in single-category EATs showed greater associations of White individuals with patriotism and being born in America, yet revealed weaker association with treating people of all races and backgrounds equally, consistent with prior findings. Further tests demonstrated a correlation between implicit bias scores and the number of images of Black individuals returned by an image ranking task, suggesting a relationship between regional prototypicality and bias. Three downstream machine learning tasks exhibited biases associating American with Whiteness, such as White individuals being identified as American in visual question answering tasks, while Asian individuals were often associated with China. In image captioning tasks, race was mentioned for Asian and Black individuals but not for White individuals. Additionally, a synthetic image generator (VQGAN) consistently lightened the skin tone of individuals of all races and generated images of White individuals with blonde hair when provided with the prompt "an American person." These findings indicate that vision-language AI models learn societal biases equating American identity with Whiteness, which subsequently propagate to downstream applications.

4 AI Bias and Transparency

While there have been attempts to address bias in word embeddings, the effectiveness of these efforts will be constrained until we gain a deeper understanding of the various intricate ways in which social biases can manifest in AI. We conducted a thorough analysis of group-based gender biases present in commonly used static English word embeddings trained on internet corpora (GloVe 2014 and fastText 2017) [Caliskan *et al.*, 2022]. We narrowed our focus to gender bias due to the consistent and noteworthy associations observed in our prior work. By examining these embeddings, we provided a comprehensive assessment of the extent of gender biases.

Among the 1,000 most frequent words in the vocabulary, 77% are more associated with men than women, providing direct evidence of a masculine default, which persists over the entire lexicon, in the everyday language of the online English-speaking world. When examining parts-of-speech, we found that the top male-associated words are typically verbs (e.g., fight, overpower) while the top female-associated words are typically adjectives and adverbs (e.g., giving, emotionally). To visually illustrate the gender association measurement of the stimulus *scientist* via SC-WEAT, we present Figure 2. *scientist* is a male-associated word with a large effect size of $d = 0.80$ in GloVe embeddings, which were trained on a web crawl corpus comprising 800 billion tokens.

Employing a bottom-up approach, we conducted cluster analyses on the top 1,000 words associated with each gender. Our findings revealed distinct patterns: the male-associated concepts predominantly encompassed roles and domains related to big tech, engineering, cars, religion, sports, and violence. Conversely, the female-associated concepts displayed female-specific slurs and sexual content, as well as appearance, relationship, and kitchen terms. In our recent research, we have developed new approaches to explore

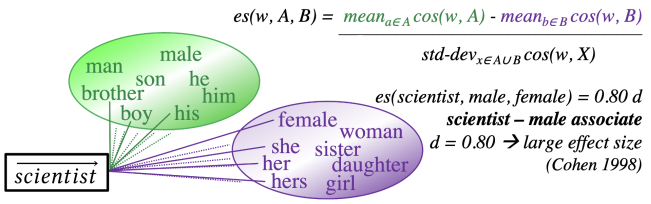

Figure 2: Illustration of SC-WEAT in measuring the effect size (d) of standardized differential association of the single stimulus (*scientist*) with two concepts (*men* and *women*).

non-binary, transgender, and fluid representations of gender within intersectional contexts [Omrani Sabbaghi *et al.*, 2023; Ghosh and Caliskan, 2023]. By employing these approaches, we expand the study of gender bias in word embeddings beyond the investigation of semantic relationships alone. Instead, we aim to examine gender differences in multiple textual manifestations, providing a more comprehensive understanding of the sources of gender associations in language.

By evaluating nine vision-language AI models trained on web scrapes using the CLIP objective, we have gathered compelling evidence supporting a bias extensively studied by psychologists: the sexual objectification of girls and women. This bias manifests when a person's human characteristics, including emotions, are disregarded, reducing them to a mere body or a collection of body parts [Wolfe *et al.*, 2023]. Through replication of three experiments established in psychology literature, we demonstrate that this phenomenon persists within trained AI models.

In the first experiment, standardized images of women from the Sexual OBjectification and EMotion Database were utilized. The results revealed a dissociation between human characteristics and objectified women in the models. The recognition of emotional states by the model was influenced by whether the subject was fully or partially clothed. Notably, Grad-CAM saliency maps for prompts that include emotion content (e.g., "a photo of a sad person") highlighted the model's tendency to be distracted from emotional expressions to partially clothed areas in images of objectified women.

The second experiment investigated the effect within an automatic image captioner, Antarctic Captions. Words denoting emotion were less frequently included in the image captions of objectified women compared to non-objectified women. In the third experiment, images of non-objectified female professionals (such as scientists, doctors, and executives) were more likely to be associated with sexual context in language when compared to images of male professionals. This phenomenon is known to impact female professionals' careers. Lastly, the fourth experiment demonstrated that a prompt mentioning "a [age] year old girl" (for ages 12 to 18) generated sexualized images (as determined by a verified NSFW classifier) in up to 73% of cases for VQGAN-CLIP, and up to 42% for Stable Diffusion. In contrast, the corresponding rate for boys never exceeded 9%. The evidence indicates that vision-language AI models trained on automatically collected web scrapes learn biases of sexual objectification, which propagate to generated outputs and downstream applications. These findings raise ethical and legal concerns.

Advances in AI enabled the conversion of user-written text descriptions into realistic images. These generative AI models are readily accessible online, leading to the generation of millions of images daily. We investigated Stable Diffusion and Dall-E and found that they amplify dangerous and complex stereotypes, which are challenging to predict and mitigate by users and developers [Bianchi *et al.*, 2023].

5 Conclusions

My research agenda focuses on evaluating associations and biases in AI that are learned from human society, analyzing how these associations propagate to AI outputs, assessing the societal impact of biases, and examining their ethical implications. Specifically, I investigate how machines perceive social groups and how they make decisions concerning social entities. The proliferation of easily accessible generative AI models presents opportunities for automated decision-making and practical human-AI interaction scenarios. However, it also gives rise to various ethical concerns, as AI models tend to perpetuate and amplify complex biases. To advance this field, I develop transparency enhancing approaches that analyze bias throughout the entire AI lifecycle. These efforts lay the foundation for responsibly developing and deploying AI while taking into account its ethical implications through empirical methodologies. Evaluating the transfer of information between human society and AI, as models are trained and humans interact with AI systems, forms an essential aspect of my research agenda in trustworthy AI.

Ethical Statement

This line of research is intrinsically intertwined with empirical ethics and draws inspiration from humanist perspectives. Its primary focus is on the quantitative study of AI, bias, and ethics, particularly within the domains of machine learning, natural language processing, and computer vision. The investigation of AI bias directly pertains to well-regulated domains in society, making this research agenda crucial for informing evidence-based technology policy. A central tenet of this research is to prioritize human rights and societal considerations while advancing AI responsibly. By developing AI bias evaluation methods and transparency enhancing approaches that are inclusive and adapt to evolving norms, harms, and risks associated with AI, we contribute to the development and deployment of trustworthy AI.

Acknowledgments

This work is supported by the U.S. National Institute of Standards and Technology (NIST) Grant 60NANB20D212T. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect those of NIST. I am immensely grateful for the invaluable contributions made by Ph.D. students, undergraduate and graduate student researchers, collaborators, co-authors, mentors, and advisors, all of whom have played instrumental roles in advancing this research agenda. Lastly, I would like to express my gratitude to the IJCAI 2023 Program Committee for inviting me to deliver a spotlight talk in the Early Career Spotlight Track at IJCAI 2023.

References

[Bianchi *et al.*, 2023] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*, 2023.

[Caliskan and Lewis, 2021] Aylin Caliskan and Molly Lewis. Social biases in word embeddings and their relation to human cognition. *Handbook of Language Analysis in Psychology. Guilford Press*, 2021.

[Caliskan and Steed, 2022] Aylin Caliskan and Ryan Steed. Managing the risks of inevitably biased visual artificial intelligence systems. *Brookings Institution*, 2022.

[Caliskan *et al.*, 2016a] Aylin Caliskan, Joanna Bryson, and Arvind Narayanan. A story of discrimination and unfairness: Implicit bias embedded in language models. *9th Hot Topics in Privacy Enhancing Technologies (HotPETs 2016)*, Accepted on May 20, 2016. https://www.gffz.de/fileadmin/user_upload/LAKOF/informatik/A_Story_of_Discrimination_and_Unfairness.pdf.

[Caliskan *et al.*, 2016b] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora necessarily contain human biases. *arXiv preprint arXiv:1608.07187*, pages 1–14, 2016.

[Caliskan *et al.*, 2017] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[Caliskan *et al.*, 2022] Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R. Banaji. 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 (AAAI/ACM AIES)*, 2022.

[Caliskan, 2021] Aylin Caliskan. Detecting and mitigating bias in natural language processing. *Brookings Institution*, 2021.

[Charlesworth *et al.*, 2022] Tessa Charlesworth, Aylin Caliskan, and Mahzarin R. Banaji. Historical representations of social groups across 200 years of word embeddings from google books. *Proceedings of the National Academy of Sciences (PNAS)*, 2022.

[Ghosh and Caliskan, 2023] Sourojit Ghosh and Aylin Caliskan. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AAAI/ACM AIES)*, 2023.

[Greenwald and Banaji, 1995] Anthony G Greenwald and Mahzarin R Banaji. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1):4, 1995.

[Greenwald *et al.*, 1998] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464, 1998.

[Guo and Caliskan, 2021] Wei Guo and Aylin Caliskan. 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 (AAAI/ACM AIES)*, pages 122–133, 2021.

[Mei *et al.*, 2023] Katelyn X. Mei, Sonia Fereidooni, and Aylin Caliskan. Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*, 2023.

[Omrani Sabbaghi and Caliskan, 2022] Shiva Omrani Sabbaghi and Aylin Caliskan. Measuring gender bias in word embeddings of gendered languages requires disentangling grammatical gender signals. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AAAI/ACM AIES)*, 2022.

[Omrani Sabbaghi *et al.*, 2023] Shiva Omrani Sabbaghi, Robert Wolfe, and Aylin Caliskan. Evaluating biased attitude associations of language models in an intersectional context. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AAAI/ACM AIES)*, 2023.

[Pandey and Caliskan, 2021] Akshat Pandey and Aylin Caliskan. Disparate impact of artificial intelligence bias in ridehailing economy's price discrimination algorithms. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AAAI/ACM AIES)*, pages 822–833, 2021.

[Steed and Caliskan, 2021a] Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pretraining contain human-like biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*, pages 701–713, 2021.

[Steed and Caliskan, 2021b] Ryan Steed and Aylin Caliskan. A set of distinct facial traits learned by machines is not predictive of appearance bias in the wild. *AI and Ethics*, 1(3):249–260, 2021.

[Toney *et al.*, 2021] Autumn Toney, Akshat Pandey, Wei Guo, David Broniatowski, and Aylin Caliskan. Automatically characterizing targeted information operations through biases present in discourse on Twitter. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 82–83. IEEE, 2021.

[Toney-Wails and Caliskan, 2021] Autumn Toney-Wails and Aylin Caliskan. Valnorm quantifies semantics to reveal consistent valence biases across languages and over centuries. *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[Wolfe and Caliskan, 2021] Robert Wolfe and Aylin Caliskan. Low frequency names exhibit bias and over-

fitting in contextualizing language models. *Proceedings of the 2021 conference on empirical methods in natural language processing (EMNLP)*, 2021.

[Wolfe and Caliskan, 2022a] Robert Wolfe and Aylin Caliskan. American==white in multimodal language-and-image ai. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AAAI/ACM AIES)*, 2022.

[Wolfe and Caliskan, 2022b] Robert Wolfe and Aylin Caliskan. Contrastive visual semantic pretraining magnifies the semantics of natural language representations. *Association for Computational Linguistics (ACL)*, 2022.

[Wolfe and Caliskan, 2022c] Robert Wolfe and Aylin Caliskan. Detecting emerging associations and behaviors using regional and diachronic word embeddings. In *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*. IEEE, 2022.

[Wolfe and Caliskan, 2022d] Robert Wolfe and Aylin Caliskan. Markedness in visual semantic ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*, 2022.

[Wolfe and Caliskan, 2022e] Robert Wolfe and Aylin Caliskan. Vast: The valence-assessing semantics test for contextualizing language models. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[Wolfe *et al.*, 2022] Robert Wolfe, Mahzarin R. Banaji, and Aylin Caliskan. Evidence for hypodescent in visual semantic ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*, 2022.

[Wolfe *et al.*, 2023] Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. Contrastive language-vision ai models pretrained on web-scraped multimodal data exhibit sexual objectification bias. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*, 2023.