

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Geometric Artifacts or Semantic Bias? The Effect of Embedding Space Anisotropy on Implicit Bias Measurement

Permalink

<https://escholarship.org/uc/item/3bq0c5fw>

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Ono, Seitaro

Saiki, Jun

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Geometric Artifacts or Semantic Bias? The Effect of Embedding Space Anisotropy on Implicit Bias Measurement

Seitaro Ono
Kyoto University, Kyoto, Japan

Jun Saiki
Kyoto University, Kyoto, Japan

Abstract

The Word Embedding Association Test (WEAT) was originally developed to assess implicit biases in language models and has since been widely employed as a tool for investigating implicit social biases in human cognition. However, recent research has revealed "anisotropy" in embedding spaces, a phenomenon whereby vectors concentrate in narrow regions, raising concerns that such geometric distortions may confound measurements of implicit social bias. This study examines whether WEAT scores reflect genuine semantic associations or geometric artifacts. We applied a whitening transformation to correct for spatial anisotropy and compared bias scores before and after correction across multiple WEAT categories. Results demonstrated that bias scores decreased substantially following geometric correction, with one racial bias category being eliminated. These findings suggest that standard bias measurements may conflict geometric distortions with semantic content, highlighting the necessity of spatial correction when using embedding-based methods to study human implicit cognition.