American == White in Multimodal Language-and-Image AI

Robert Wolfe
University of Washington
Information School
Seattle, WA, USA
rwolfe3@uw.edu

Aylin Caliskan
University of Washington
Information School
Seattle, WA, USA
aylin@uw.edu

ABSTRACT

Three state-of-the-art language-and-image AI models, CLIP, SLIP, and BLIP, are evaluated for evidence of a bias previously observed in social and experimental psychology: equating American identity with being White. Embedding association tests (EATs) using standardized images of self-identified Asian, Black, Latina/o, and White individuals from the Chicago Face Database (CFD) reveal that White individuals are more associated with collective in-group words than are Asian, Black, or Latina/o individuals, with effect sizes \( > .4 \) for White vs. Asian comparisons across all models. In assessments of three core aspects of American identity reported by social psychologists, single-category EATs reveal that images of White individuals are more associated with patriotism and with being born in America, but that, consistent with prior findings in psychology, White individuals are associated with being less likely to treat people of all races and backgrounds equally. Additional tests reveal that the number of images of Black individuals returned by an image ranking task is more strongly correlated with state-level implicit bias scores for White individuals (Pearson’s \( \rho = .63 \) in CLIP, \( \rho = .69 \) in BLIP) than are state demographics (\( \rho = .60 \)), suggesting a relationship between regional prototypicality and implicit bias. Three downstream machine learning tasks demonstrate biases associating American with White. In a visual question answering task using BLIP, 97% of White individuals are identified as American (Pearson’s \( \rho = .60 \) in CLIP, \( \rho = .63 \) in BLIP) than are state demographics (\( \rho = .24 \)), suggesting a relationship between regional prototypicality and implicit bias. Three downstream machine learning tasks demonstrate biases associating American with White. In a visual question answering task using BLIP, 97% of White individuals are identified as American (Pearson’s \( \rho = .60 \) in CLIP, \( \rho = .63 \) in BLIP) than are state demographics (\( \rho = .24 \)), suggesting a relationship between regional prototypicality and implicit bias.

CSS CONCEPTS

- Computing methodologies → Computer vision; Natural language processing.

KEYWORDS

bias in AI, multimodal models, visual semantics, racial bias

ACM Reference Format:

Robert Wolfe and Aylin Caliskan. 2022. American == White in Multimodal Language-and-Image AI. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES’22), August 1–3, 2022, Oxford, United Kingdom. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3514094.3534136

1 INTRODUCTION

The United States is a multiethnic and pluralistic country whose citizens espouse a commitment to the principles of equality and inclusion in the American identity for all people, regardless of race or ethnicity [23, 58]. Yet, as in many societies which outwardly express a commitment to the ideals of equality, the structure of American society distributes opportunities unequally [42], and experimental psychologists have found that, even among Americans who affirm a belief in equality for all races, to be American is implicitly associated with being White [15].

Artificial Intelligence (AI) is increasingly used to determine access to the benefits, responsibilities, and opportunities which attend life as an American. Computer vision, in particular, is employed to monitor and police American territorial borders and ports of entry [46], and but for bipartisan public outcry [51] informed by the research of Buolamwini and Gebru [5], the U.S. Internal Revenue Service (IRS) would have required American citizens to use third-party facial recognition AI to create an online IRS account [59]. While most deployed systems currently use supervised computer vision models designed to recognize a predefined set of image classes, the field of computer vision underwent a transformation in early 2021 with the introduction of CLIP (“Contrastive Language Image Pretraining”) [47]. CLIP is the first practical “zero-shot” language-and-image model, a system which learns to match images to descriptive text, and which performs competitively with state-of-the-art supervised models without ever explicitly training on computer vision evaluation datasets [47]. CLIP and its successors allow for the definition of image classes in natural language, and have been adapted for numerous domains, including analysis of satellite imagery [47], zero-shot object detection [24], and text-based guidance of synthetic image generators [41, 50]. Yet natural language supervision renders computer vision susceptible not only to visual biases, but to biases in human language: Goh et al. [21] find that neurons in the CLIP image encoder associate Muslims with terrorism, while Wang et al. [69] show that images of men are over-represented in occupational search queries when using CLIP for image retrieval.

The advances realized by CLIP offer an opportunity to study the propagation to AI of a bias previously observed in experimental psychology: specifically, bias associating American identity with being White [15]. Motivated by the prior discovery of implicit biases...
Bias for mediating the experiences and opportunities afforded to a joint vision-and-language embedding space. The broader class of models referred to as language-and-image AI. Form language and image representations and may use them in downstream tasks are adapted from the work of Devos and Banaji [15]. The images most associated with a text prompt "a photo of someone who lives in the state of [state]" are returned for each of the fifty U.S. states and the District of Columbia. Results correspond to statistics collected in the 2019 U.S. census, with R2 coefficient of up to .74 for CLIP, .54 for BLIP, and .33 for SLIP, corresponding to the size of each model’s training data (400 million, 129 million, and 15 million image-text pairs, respectively). However, language-and-image AI tends to overestimate the population of Black, Asian, or Latina/o individuals living in a state, indicating that location-based categorical associations in AI are not the straightforward result of “learning” the population of a state, but reflect biases expressed in human descriptions. Results also reveal that the proportion of images of Black individuals returned correlates positively with the implicit racial bias of White individuals living in a state, and that the correlation is stronger than correlations between human implicit bias and population statistics. This suggests that implicit bias is related to the threat to the prototypicality of White individuals, as measured via representation in AI.

(1) Visual semantic embedding spaces exhibit correspondence with the demographics of U.S. states, with R2 coefficient of up to .74, in CLIP. An image ranking task uses the photos of the Chicago Face Database (CFD), a database of standardized images of self-identified Asian, Black, Latina/o, and White individuals intended for the psychologi- cal study of race [37]. The images most associated with the text prompt “a photo of someone who lives in the state of [state]” are returned for each of the fifty U.S. states and the District of Columbia. Results correspond to statistics collected in the 2019 U.S. census, with R2 coefficient of .74 for CLIP, .54 for BLIP, and .33 for SLIP, corresponding to the size of each model’s training data (400 million, 129 million, and 15 million image-text pairs, respectively). However, language-and-image AI tends to overestimate the population of Black, Asian, or Latina/o individuals living in a state, indicating that location-based categorical associations in AI are not the straightforward result of “learning” the population of a state, but reflect biases expressed in human descriptions. Results also reveal that the proportion of images of Black individuals returned correlates positively with the implicit racial bias of White individuals living in a state, and that the correlation is stronger than correlations between human implicit bias and population statistics. This suggests that implicit bias is related to the threat to the prototypicality of White individuals, as measured via representation in AI.

(2) Visual semantic embedding spaces associate American with White. Two language-image association tests are adapted from the work of Devos and Banaji [15]. The first is a We vs. They embedding association test (EAT), testing association with collective in-group words vs. out-group words.

When the term “visual semantic” is used in this research, it refers specifically to a joint vision-and-language embedding space. The broader class of models which form language and image representations and may use them in downstream tasks are referred to as language-and-image AI.

CFD images of people who self-identify as Asian, Black, Latina/o, and White are the target groups of the EAT, for which the differential association is tested. Across three models, images of White individuals are consistently associated with collective in-group words when compared with images of Asian, Black, or Latina/o individuals, with all effect sizes for the White vs. Asian comparison > 0.4. A second EAT measures association with three characteristics found by Devos and Banaji [15] to best describe what American identity according to study participants: patriotism, nativism (being born in the U.S.), and egalitarianism (in this case, treating people of all races and backgrounds equally). Across models, White individuals are differentially associated with patriotism and with nativism. However, White individuals are strongly disassociated from egalitarianism. These findings in AI reflect those of Devos and Banaji [15].

(3) Language-and-Image AI associates American identity with being White in three downstream machine learning tasks: visual question answering, image captioning, and text-guided synthetic image generation. Prompted with the question “Is this person an American?”, the BLIP visual question answering head answers “Yes” 97% of the time for White individuals, 56% of the time for Black individuals, 61% of the time for Latina/o individuals, and only 3% of the time for Asian individuals. When prompted with the question “What state does this person live in?”, the BLIP visual question answering head answers “China” 53% of the time for Asian individuals, while responding with the name of a U.S. state more than 97% of the time for all other images. Prompted to automatically generate a caption for each photo in the Chicago Face Database, the BLIP image captioning head never remarks on the race or ethnicity of White individuals, but remarks on the race or ethnicity of Asian individuals up to 36% of the time, and up to 18% of the time for Black individuals. The race or ethnicity of White individuals is never described. Prompted to generate an image of “an American person,” a synthetic image generator guided by the features of CLIP produces images of White individuals with blonde hair. As seen in Figure 1, when provided with an initialization image of an individual from the CFD, the CLIP-guided image generator changes the color of the individual’s skin to white. For initialization images of Black individuals, mean pixel brightness for a 50x50 pixel crop of the forehead increases by 35% after 80 training iterations, reflecting lightening of skin tone.
The present research indicates that language-and-image AI associates American with White in ways that reflect human biases through similar mechanisms of association. This research also indicates that biases of association with American identity observed in the embedding spaces of language-and-image models influence biases in downstream applications of vision and language AI.

2 RELATED WORK

Relevant work concerning associations of American with being White, racial bias in AI, and language-and-image AI is reviewed.

American Identity Bias This research is grounded in foundational work in experimental psychology by Devos and Banaji [15], who observe that, despite explicit affirmation from study participants that members of different ethnic groups should be treated equally, the category ‘American’ is implicitly synonymous with being White for American subjects. The six studies demonstrating this bias included Implicit Association Tests (IATs), which measure unconscious or "implicit" biases and associations in human subjects [22] showing that participants more easily paired White faces with American symbols and landmarks than they did Asian faces, and IATs showing that participants more easily paired the names of White Europeans (not Americans) with American symbols than they did the names of African Americans and Asian Americans [15]. Subsequent research affirms that a societal in-group may make itself synonymous with a larger human category. Wenzel et al. [70] find that members of an ethnocentric in-group tend to generalize their characteristics onto a positively valued superordinate category to increase the prototypicality of the in-group, and that out-group differences from the prototypical in-group norm are evaluated as negative deviations. Danbold and Huo [12] find that fear of losing such prototypical status in America predicts White Americans’ greater support for assimilation, and lower support for diversity.

Other psychological research has sought to model the interaction of racial status and perceived foreignness in the U.S. and Cheryan [80] surveyed Asian, Black, Latina/o, and White American subjects regarding their recent and overall experiences with racial prejudice in the U.S., and model U.S. racial dynamics using a quadrant of perceived superiority vs. inferiority and perceived American identity vs. foreignness, wherein White Americans are perceived and treated as superior and American; Black Americans are perceived and treated as inferior and relatively American; Latina/o Americans are perceived and treated as inferior and foreign; and Asian Americans are perceived and treated as foreign but relatively superior to Black and Latina/o Americans. Recent work finds that ethnocentric biases may change along with regional demographics: Devos et al. [16] found that steeper linear increases in the proportion of Asian Americans living in a metropolitan area over time are associated with greater implicit inclusion in national identity.

Research in psychology also finds that ethnocentric American biases have consequences for the mental health and opportunities afforded to Asian Americans and Latina/o Americans. Armenta et al. [2] found that, among U.S.-born Asian Americans and Latina/o Americans, perceived objectification as a foreigner correlates with lower life satisfaction, greater depressive symptoms, and lower self-esteem. Huyhn et al. [28] find that awareness of the perpetual foreigner stereotype, which asserts that people who are not White will always be seen as foreign in the U.S., predicts lower sense of belonging to American culture, lower hope and life satisfaction for Asian Americans, and greater incidence of depression for Latina/o Americans. Yogeeswaran and Dasgupta [78] find that the stronger an individual’s implicit bias associating White with the prototypical American, the less willing the individual is to hire Asian Americans in national security jobs. Huyhn et al. [27] find a relationship between White prototypicality and anti-minority policy attitudes and acculturation ideologies in a study of White Americans.

The present research adapts two of the studies of Devos and Banaji [15] to examine biases in language-and-image AI. The first is a We/They IAT, wherein one group of words (we, our, ourselves) implies a collective in-group identity, while another group of words (they, other, themselves) implies out-group identity [15]. Devos and Banaji [15] find that it is easier for participants to pair in-group words with the faces of individuals with whom they share an ethnic identity. The second study of Devos and Banaji [15] adapted in this work is drawn from a survey of participants to assess the explicit attributes most associated with American identity. Devos and Banaji [15] find that three attributes are most associated with American identity: egalitarianism, or treating people of all races and backgrounds equally; patriotism; and native status, or being born in America. Of these attributes, Asian Americans were perceived as the least likely to be patriotic, and the least likely to have been born in America, while White Americans were perceived as being less egalitarian than Asian Americans or African Americans [15].

Bias in AI Prior research on racial bias in AI has addressed, primarily, four prongs: failures of generalization of state-of-the-art technology [5]; under-representation in machine learning datasets [17, 74]; biases of association reflected in embedding spaces [7, 25]; and downstream biases in applications of AI [44]. The current research examines questions of bias and identity both in embedding spaces and in downstream tasks demonstrating disparate impact.

Explicit Bias and the Word Embedding Association Test A significant component of the present research is the adaptation of methods grounded in experimental psychology to observe biases in machine learned representations. Foundational research on human-like biases in word embeddings was contributed by Caliskan et al. [7], who introduced the Word Embedding Association Test (WEAT), an adaptation of the IAT of Greenwald et al. [22] which showed that machine-learned semantics derived from internet-scale web corpora reflect the biases of the populations who produce them. The WEAT measures the differential angular similarity between two groups of target words with two groups of attribute words, and returns an effect size (Cohen’s $d$) [10] and a $p$-value indicating statistical significance. The single-category SC-WEAT measures the differential angular similarity of a single word with two attribute groups. The WEAT and SC-WEAT allow for the measurement of association with concepts, and subsequent work extends the WEAT to contextualized word embeddings [25, 77] and sentence embeddings [38] in language models such as BERT [14] and GPT-2 [48]. The appendix includes formulae for the WEAT and SC-WEAT.

Impact of Training Data The composition of machine learning training datasets has been shown to have significant impacts on the biases learned by a model. Brunet et al. [4] show that the least frequently occurring words in static word embedding training corpora are also the most biased, and have the most unstable representations.
Wolfe and Caliskan [74] find that names belonging predominantly to Asian, Black, and Latina/o Americans are underrepresented in the training corpora of language models, leading to bias and overfitting in the pretrained model. Dodge et al. [17] find that methods intended to prevent bias in constructing the C4 corpus also remove data created by and about marginalized populations. Caliskan et al. [6] show that word embeddings trained on internet-scale corpora reflect a masculine default which pervades the embedding space.

**Bias in Computer Vision** Both semantic biases and biases related to failures of generalization for underrepresented populations have been observed previously in computer vision. Buolamwini and Gebru [5] find that the underrepresentation in computer vision datasets results in facial recognition models which disproportionately fail to recognize the faces of women with darker skin. Wilson et al. [71] finds that state-of-the-art object detection systems also fail for people with darker skin. Rhue [53] observes that emotion detection systems are more likely to ascribe negative emotions to Black individuals, while Kim et al. [31] find that emotion detection systems fail to generalize for images of older adults. In accordance with this finding, Park et al. [45] show that computer vision datasets systematically underrepresent older adults. Steed and Caliskan [62] find that the embedding structure of generative image models such as Image GPT [9] is reflective of humanlike social biases, and that the model generates stereotypically sexualized images of women.

**Reflection of Human Society to AI** In introducing the SC-WEAT, Caliskan et al. [7] demonstrate a linear relationship between the SC-WEAT association of the name of a profession with female attribute words and the proportion of women employed in the profession, with Pearson’s $\rho = .88$. The veridical properties of static word embeddings have also rendered them a useful tool for studying human societies. Kozlowski et al. [32] use the geometry of static word embeddings trained on Google N-grams over decades of the twentieth century to show that the material markers used to signify social class changed with the economic transformations of the century. Walter et al. [68] use diachronic word embeddings trained on German parliamentary proceedings to study the evolution of German political biases over time. Joseph and Morgan [30] show that word embeddings can capture population-level beliefs which correspond to the results of surveys of that population. That survey results correlate with bias in embedding spaces is notable for the current research, which adapts results from a survey designed by Devos and Banaji [15] to test for biases in an embedding space.

**Language-and-Image AI** This research examines bias in three language-and-image AI models: CLIP, SLIP, and BLIP.

**Language-Image Pretraining** CLIP was the first in a new generation of multimodal language-and-image models trained using natural language supervision [47]. Where supervised computer vision trains on a defined set of image classes and associated images, CLIP instead learns to pair text captions collected from the internet with their associated images [47]. While the objective is straightforward, the results introduced a new paradigm in computer vision, wherein a user of the pretrained model can define their own image classes in natural language, and retrieve, rank, or classify images based on association with the text [47]. In this sense, CLIP is the first “zero-shot” image associator, and is not dependent on the classes and images of any specific dataset [47]. The architecture of CLIP is composed of a contextualizing language model (a smaller version of the causal language model GPT-2 [48], based on the transformer architecture of Vaswani et al. [67]) used to form sentence embeddings, and an image encoder, either a Vision Transformer [18] or a ResNet [26]. The language and vision models are jointly pretrained, and the representations formed by each are projected into a joint “visual semantic” embedding space [47], wherein cosine similarity is used to assess the similarity between embedded image and embedded text. This research examines the CLIP-ViT-Base-Patch32 model available via the Transformers library of Wolf et al. [72], the most downloaded model of the versions available via Transformers.

SLIP adopts the architecture and training design of CLIP and adds data augmentation to the CLIP objective [39]. SLIP randomly resizes and crops images such that they are between 50% and 100% of the size of the original image, a technique intended to improve the robustness of the model for extracting the semantic content of an image [39]. SLIP also adds an image-based self-supervision branch, wherein the model is trained to represent different views of the same image with similar vectors [39]. This research examines the ViT-Base version of SLIP trained for 100 epochs, the best performing version of the Base model on zero-shot ImageNet evaluation [39].

BLIP is a multimodal language-and-image encoder-decoder model [35]. Like CLIP and SLIP, BLIP trains a visual semantic embedding space using contrastive loss to align text and image representations [35]. Unlike CLIP and SLIP, BLIP is also trained for language-and-image tasks which require text generation, including visual question answering and image captioning, on which it set new state of the art [35]. BLIP introduced a new synthetic caption generation and filtering technique, known as CapFilt, which filters out noisy or uninformative captions during training [35]. This research uses the ViT-Base version of BLIP trained on 129 million image-text pairs with MS-COCO fine-tuning for evaluating embedding space associations; the ViT-Base checkpoint with CapFilt-L for automatic image captioning; and the ViT-Base checkpoint with CapFilt-L for visual question answering [35]. These are the default checkpoints used in a publicly available version of the model [35].

**Visual Semantic AI for Guiding and Training Generative Models** Among the first uses of CLIP was to train the text-to-image generation model DALL-E [50], and CLIP subsequently used in the training of GLIDE [41] and DALL-E 2 [49], diffusion-based text-to-image generation models. Such models use CLIP representations as ground truth, and are likely to inherit its biases. Because no version of DALL-E or GLIDE capable of generating human images is publicly available, this research examines a CLIP-guided VQGAN [19], which uses convolutional neural networks to learn a vocabulary of image components, and transformers to compose them in high-resolution images [19]. VQGAN-CLIP uses the cosine similarity between CLIP-embedded text and VQGAN-generated images as a loss to increase the image’s similarity with the target text [11].

**Bias Specific to Language-and-Image AI** Radford et al. [47] and Agarwal et al. [1] find that CLIP prefers text highlighting the physical features of women, and is more likely to misclassify Black individuals into animal categories. Wolfe et al. [73] provide evidence that CLIP associates images of multiracial individuals with race or ethnicity labels according to a rule of hypodescent, or one-drop rule. Wang et al. [69] mitigate gender biases in CLIP image retrieval results by muting gendered neurons in CLIP image embeddings, and by preferentially sampling images of women. Wolfe and Caliskan
[76] find that CLIP draws attention to the race, gender, and age of underrepresented individuals, while leaving these characteristics unmarked for white, male, and middle-aged individuals.

3 DATA

This research uses the Chicago Face Database (CFD) to evaluate the association of American identity with being White. The training datasets for CLIP, SLIP, and BLIP are also discussed.

The Chicago Face Database The CFD is a dataset of images used to study race and ethnicity in psychology [37]. The CFD includes 597 high-resolution (2, 444 x 1, 718 pixel) images of male and female volunteers who provided information regarding their self-identified race or ethnicity [37]. The self-identified races and ethnicities reported for the CFD include Asian, Black, Latina/o, and White [37]. CFD images position subjects facing the camera against a white background, such that each subject’s face occupies the same area of the image. All subjects are captured with a “neutral” facial expression, and a subset with “happy (open mouth),” “happy (closed mouth),” “angry,” and “fearful” facial expressions. In accordance with the methodology of Devos and Banaji [15], this research uses only images of subjects with a neutral facial expression. Because the present research examines biases related to American identity, it is noteworthy that all CFD subjects were recruited in the U.S.

U.S. Census Data Results for an image ranking experiment are compared to 2019 U.S. state-level census population estimates available at https://www.census.gov/data/tables/time-series/demo/popest/20190s-state-detail.html.

IAT Data This research examines the correlation of the racial association of U.S. states in language-and-image AI with the mean IAT effect sizes for those states. The IAT data for this analysis is obtained from the Project Implicit and is based on eight years of state-level data obtained via the online IAT [52].

Training Data for Language-and-Image AI The training data for AI models determines to a large extent the biases learned [7, 74]. The training datasets for CLIP, SLIP, and BLIP are discussed below.

CLIP Radford et al. [47] train CLIP on the WebImageText corpus (WIT), a web scrape composed of 400 million images and associated captions. Radford et al. [47] produce the query list using every bigram from Wikipedia with high pointwise mutual information, the names of Wikipedia articles, and all WordNet synsets [47].

SLIP SLIP’s training data is drawn from an English-language subset of the Yahoo Flickr Creative Commons (YFCC100M) dataset [63], and includes 15-million image-text pairs. YFCC100M includes more than 11 million images of people posted to the internet between 2002 and 2014 [63]. Note that while SLIP outperforms the CLIP architecture when both models are trained on the 15-million image-text dataset, SLIP does not outperform the versions of CLIP trained on 400 million images [39]. This research evaluates a version of CLIP trained on 400 million images, but only has access to a version of SLIP trained on 15 million images; thus, results obtained from CLIP are expected to better reflect societal associations and biases.

BLIP BLIP’s training dataset consists of 14-million image-text pairs drawn from five datasets: COCO (“Common Objects in Context”), an in-context object detection dataset [36]; Visual Genome, a densely annotated language-image dataset to enable recognition of relationships between objects in images [33]; Conceptual Captions, a language-image dataset designed to train automatic image captioning AI [60]; Conceptual 12M, a language-image dataset which relaxes the data collection pipeline of Conceptual Captions to include more data for language-image training [8]; and SBU captions, a collection of over a million images and captions collected and filtered from Flickr [43]. Most BLIP models, including those examined in this research, also train on a subset of 115-million image-text pairs from the LAION-400M open source dataset, which is intended to imitate the WIT dataset [47], and which uses CLIP cosine similarity measurements to filter low-quality image-text pairs [37]. Birhane et al. [3] found evidence of pornographic, misogynistic, and stereotypical images and text in LAION-400m [56].

4 APPROACH AND EXPERIMENTS

Three experiments evaluate the bias associating American identity with being White in multimodal language-and-image AI. First, an experiment tests the ability of multimodal spaces to encode biases and veridical demographic information using an image ranking algorithm. Second, adaptations of the WEAT and SC-WEAT are used to evaluate bias related to American identity in visual semantic embedding spaces. Third, bias related to American identity is assessed in three downstream language-and-image tasks: visual question answering, image captioning, and synthetic image generation.

State-Level Correlations with Bias and Demographics The correspondence of visual semantic representations with U.S. state-level demographics is assessed using an image ranking algorithm. Each of the 597 images of the CFD are embedded using the vision transformer of the multimodal model in question. The image embeddings are randomly balanced based on race or ethnicity to account for group size imbalances, such that each of the four races or ethnicities in the CFD has 108 embedded images, for 432 total. Then, an embedding reflecting the name of each state is obtained by providing the linguistic context “a photo of someone who lives in [STATE]”, based on the prompting format suggested by Radford et al. [47] and adopted by Mu et al. [39], to the text encoder of the multimodal model. The cosine similarity of the text embedding for each state is obtained with all 432 image embeddings, and the images are ranked by cosine similarity from greatest to least. The 108 highest cosine similarities are selected, and the number of images corresponding to each race or ethnicity in the CFD are counted. Using the 108 highest cosine similarities allows a state to be entirely associated with a single race or ethnicity, if this is what the model returns. To adjust for the effects of randomly downsampling, this process is repeated 1,000 times for each state, and the mean count of images returned for each race or ethnicity is obtained. Results are evaluated based on Pearson’s $\rho$ of the percent of images returned and the percent population of each state for each race or ethnicity. Correspondence between census statistics and images returned for all four races or ethnicities are obtained by fitting a multivariate linear regression and reporting the $R^2$ coefficient.

After measuring correspondence with state demographics, the number of images of Black individuals for each state is evaluated against state-level measurements of implicit bias for White online
IAT participants and for Black online IAT participants. The intention of this experiment is to quantify whether the prototypical racial association of a state predicts the biases of White and Black individuals living in that state, as in research finding that fear of losing prototypical status predicts bias in White Americans [12].

**EAT and SC-EAT** The present research employs a language-image version of the WEAT, which will be referred to as the EAT (Embedding Association Test), because the test is not restricted to the modality of language, as in the WEAT. An EAT and three SC-EATs are used to evaluate biases in the visual semantic embedding spaces of CLIP, SLIP, and BLIP. As defined by Caliskan et al. [7], EAT and SC-EAT approaches require attribute groups A and B to be of equal size, and the EAT requires target groups X and Y to be of the same size. However, this research uses the images of the CFD in attribute and target groups, and the number of images included in the CFD for each race or ethnicity is unequal. When population sizes are unequal, Cohen’s $d$ can be obtained using a pooled standard deviation, for which the formula is provided in the appendix. $p$-value is obtained using a Welch’s $t$-test, which does not assume equal variance or population size. Cohen [10] defined an effect size of .2 as small, .5 as medium, and .8 as large.

**We/They EAT** In evaluating bias related to American identity, Devos and Banaji [15] design IAT attribute groups to represent prototypical in-group status and out-group status. One such test involves a task which pairs faces of White individuals with a ‘We’ attribute group and faces of Asian individuals with a ‘They’ attribute group. The stimuli used by Devos and Banaji [15] are as follows:

- **We:** we, our, ourselves
- **They:** they, other, themselves

Unlike the IAT, the EAT requires sets of at least 8 word or image stimuli to ensure the statistical significance of the results. Thus, the We/They attribute word groups are expanded to include 8 words per group, which are close variations of the originals:

- **We:** we, our, ourselves, ours, us, familiar, similar, here
- **They:** they, their, themselves, theirs, them, other, others, there

Words referring to an individual, such as ‘I’ and ‘me,’ are not included in the “We” group, as Devos and Banaji [15] note that the attribute words are intended to represent a collective national identity and a collective other. For a language-image EAT, the We/They groups are attribute groups A and B in the EAT formula. The target group X is a collection of image embeddings of White individuals in the CFD. The target group Y is a collection of image embeddings of Asian individuals, Black individuals, or Latina/o individuals in the CFD. For this test, a positive effect size indicates greater similarity of in-group words with the White target group, and a negative effect size indicates greater similarity of in-group words with a Black, Asian, or Latino/a target group.

**American Trait SC-EAT** Devos and Banaji [15] survey study participants to identify the core traits connected with American identity. The results of the survey indicate that three traits are more central than any others surveyed: whether a person is patriotic; whether they are native to the U.S.; and whether they treat people of all races and backgrounds equally, commensurate with the explicit ideology of racial equality in the U.S. Association of these three traits with the races and ethnicities included in the CFD is tested using the SC-EAT, with the following three phrases used as the target embedding:

- "a photo of someone who is patriotic"
- "a photo of someone who is an immigrant to America"
- "a photo of someone who treats people of all races and backgrounds equally"

For these tests, the attribute group A is composed of image embeddings of White individuals and the attribute group B is composed of image embeddings of Asian, Black, or Latina/o individuals. A positive effect size indicates association of the phrase with White individuals, while a negative effect size indicates association of the phrase with Asian, Black, or Latina/o individuals. These tests observe the “a photo of” prompting format described by Radford et al. [47] and used to obtain results which correspond well to census ground truth in the experiment described previously.

**Downstream Language-and-Image Tasks** This research evaluates the presence of biases related to American identity in three tasks for which language-and-image AI is widely used: visual question answering, image captioning, and synthetic image generation.

**Visual Question Answering** The BLIP visual question answering model is prompted with a text question: “Is this person an American?” Each of the images in the CFD is input to the model, which responds with an answer as to whether the person is an American. The number of times the model answers “yes” or “no” are quantified for each race or ethnicity in the CFD, and the percentage of the time the response was “yes” is calculated. While the BLIP visual question answering model is capable of generating answers other than “yes” and “no,” and sometimes generates text such as “I don’t know,” the model responds either “yes” or “no” in all cases for this experiment. The BLIP visual question answering model is then provided with a second text question: “What state does this person live in?” Each of the images in the CFD is input to the model, and in all cases the model responds with a one-word answer. For each race or ethnicity, the answers are counted, and the percentage of the time an answer occurs for each race or ethnicity in the CFD is calculated.

**Image Captioning** Each of the images in the CFD is input to the BLIP image captioning model, and captions are generated with the top-$p$ parameter set to .5, .6, .7, .8, and .9. Top-$p$ or “nucleus” sampling controls the randomness of the output of a text generator by selecting the next token from among the minimum number of tokens which make up $p = 100\%$ of the probability mass for that word. Higher values of $p$ allow the model to select from a wider variety of sentence continuations, resulting in more varied output. At low values of $p$, such as below .5, the model is restricted to choosing from among a small subset of words, and in almost all cases generates a caption similar to “a photo of a man wearing a grey shirt” for the images of the CFD. Letting top-$p$ vary between .5 and .9 allows assessment of a wider variety of high-probability model outputs. At each level of the top-$p$ parameter, an automatically generated image is obtained for each of the CFD images. The percentage of the time a caption describes the race or ethnicity of the person in an image is calculated for each of the races or ethnicities in the CFD. Similar to the We/They EAT, this task serves as a way of measuring what races or ethnicities are prototypical, and not in need of description; and what races or ethnicities are other, and in need of description.

**Synthetic Image Generation** The VQGAN-CLIP synthetic image generation model is provided with an initialization image, which is used as the starting point for the production of an output image. Every image in the CFD is used as an initialization image. The model is provided the text prompt “an American person” to guide
the generation of a synthetic image. This experiment is performed using a checkpoint for WikiArt-16034, which is trained to generate synthetic artwork. While checkpoints exist for generating more photographic representations of faces, these checkpoints often generate unrealistic images missing facial features such as eyes or nose. This appears to be much less common with the WikiArt checkpoint.

For each of the generated images, a 50x50 pixel square of the image is cropped from the part of the face corresponding to the forehead just above the eyebrows and between the eyes. The mean brightness of the pixels in this square are measured as a proxy to understanding biases related to skin tone. This experiment assesses whether the bias associating American with White is strong enough that the model generates an image of a White individual even when provided with an initialization of a person who was recruited in America and who self-identifies as Asian, Black, or Latina/o. All synthetic images are trained for 200 iterations, the default setting of the model, and generated images are 592x592 pixels. The appendix includes additional notes concerning the design of this experiment.

5 RESULTS

Results indicate that language-and-vision AI associates American with White, and that this bias manifests in downstream tasks.

| Model  | Correlation of Retrieved Images with Census Population |
|--------|------------------------------------------------------|
|        | Asian | Black | Latina/o | White | All |
| BLIP   | 0.72  | 0.70  | 0.58    | 0.71  | 0.54 |
| CLIP   | 0.75  | 0.90  | 0.81    | 0.84  | 0.74 |
| SLIP   | 0.43  | 0.42  | 0.57    | 0.53  | 0.34 |

Table 1: The embedding spaces of BLIP, CLIP, and SLIP associate U.S. states with images of individuals who belong to the races and ethnicities living in that state, according to census demographics. CLIP trains on more than 3 times as much data as the other models, and exhibits the strongest correlation between demographics and model associations, with $R^2 = 0.74$. Correlations are Pearson’s $\rho$.

State-Level Correlations of Bias and Demographics Results reported in Table 1 indicate that strong correlations exist between the race or ethnicity of the highest ranked images for each state returned by BLIP, CLIP, and SLIP, and the demographics of that state as reported in 2019 U.S. Census data. Specifically, the $R^2$ coefficient obtained for CLIP is 0.74, for BLIP is 0.54, and for SLIP is 0.33. This corresponds to the amount of data on which the three models train, as CLIP trains on 400 million image-text pairs, BLIP on 129 million image-text pairs, and SLIP on 15 million image-text pairs. Pearson’s $\rho$ for individual races or ethnicities ranges between 0.42 and 0.90 with CLIP again achieving the highest correlations of any model, and SLIP the lowest. While these results indicate that real-world population correlates with the model’s associations, a closer look at the data reveals that images of Asian and Black individuals are generally over-represented in image ranking results relative to census population, while White individuals are under-represented relative to census population. This suggests that biases associating White with American are not the straightforward result of White individuals being the most populous U.S. racial group, but also reflect attitudes related to identity, as this research shows.

For CLIP and BLIP, the association of a state with images of Black individuals correlates positively with the mean racial bias IAT score for White individuals residing in each state. While state demographics are also positively correlated, the results show that IAT scores for White individuals correlate more strongly with the percentage of images of Black individuals returned by language-and-image AI (Pearson’s $\rho = 0.64$ in CLIP, and $\rho = 0.69$ in BLIP) than with the proportion of state population which is Black according to census figures ($\rho = 0.60$). This suggests a possible link between the degree to which the model associates a region with Black individuals, and the bias of White individuals who live in that region. Conversely, the percentage of images of Black individuals returned correlates negatively ($\rho = -0.55$ in CLIP and $\rho = -0.41$ in BLIP) with pro-White implicit bias scores for Black individuals. Figure 2 visualizes the relationship between census statistics, CLIP biases, and IAT scores.

EAT and SC-EAT Associations Findings of bias using the EAT reflect those reported in human subjects by Devos and Banaji [15]. White individuals are differentially associated with in-group words and with patriotism, while Asian, Black, and Latina/o individuals are differentially associated with egalitarianism and with not being native to America when compared with White individuals.

We/They EAT Table 2 shows that in BLIP, CLIP, and SLIP, a We/They EAT associates White individuals with in-group or collective ‘We’ words, and associates Asian individuals with out-group or ‘Other’ words, with effect sizes ranging between 0.46 and 0.64, and $\rho$-values of $10^{-3}$ or smaller. In BLIP and SLIP, White is associated with the We group and Black with the They group, while in CLIP there is no

Figure 2: The number of images returned by CLIP (center image) when prompted with “a photo of someone who lives in [STATE]” correlates strongly with 2019 census figures (left image, same scale as center image, 0-65%). However, CLIP overrepresents Black individuals relative to census statistics. Comparing to mean IAT scores for White individuals from each state (right image, min-max scaled) reveals that IAT scores correlate more strongly with national racial associations captured by AI ($\rho = 0.64$ in CLIP, $\rho = 0.69$ in BLIP) than with census statistics ($\rho = 0.60$).
Egalitarianism is differentially associated with Asian, Black, or Latina/o individuals, with statistically significant effect sizes ($\eta^2$). Only one effect size is not statistically significant, for the White vs. Black SC-EAT. In BLIP and CLIP, White is associated with the We group and Latina/o with the They group, while SLIP appears to reverse this.

### Table 2: We/They EAT

| Model | White/Asian | White/Black | White/Latina/o |
|-------|-------------|-------------|---------------|
| BLIP  | $0.51 \times 10^{-5}$ | $0.34 \times 10^{-5}$ | $0.65 \times 10^{-7}$ |
| CLIP  | $0.46 \times 10^{-3}$ | $0.51 \times 10^{-7}$ | $0.45 \times 10^{-3}$ |
| SLIP  | $0.64 \times 10^{-8}$ | $0.49 \times 10^{-9}$ | $0.49 \times 10^{-4}$ |

### Table 3: Patriotism SC-EAT

| Model | White/Asian | White/Black | White/Latina/o |
|-------|-------------|-------------|---------------|
| BLIP  | $0.56 \times 10^{-6}$ | $0.32 \times 10^{-5}$ | $0.29 \times 0.01$ |
| CLIP  | $0.28 \times 0.05$ | $0.52 \times 10^{-7}$ | $0.62 \times 10^{-7}$ |
| SLIP  | $0.35 \times 1.23 \times 10^{-28}$ | $0.13 \times 0.01$ |

### Table 4: Egalitarianism SC-EAT

| Model | White/Asian | White/Black | White/Latina/o |
|-------|-------------|-------------|---------------|
| BLIP  | $-0.97 \times 10^{-30}$ | $-3.10 \times 10^{-30}$ | $-0.09 \times 10^{-30}$ |
| CLIP  | $-1.31 \times 10^{-30}$ | $-1.96 \times 10^{-30}$ | $-0.37 \times 0.01$ |
| SLIP  | $0.84 \times 10^{-12}$ | $1.05 \times 10^{-22}$ | $0.05 \times 0.01$ |

### Table 5: The phrase "a photo of someone who is an immigrant to America" is differentially associated with Asian and Latina/o when compared with White individuals in BLIP, CLIP, and SLIP. Effect sizes are small for a White vs. Black SC-EAT in CLIP and SLIP, reflecting findings that Black individuals are seen as native to the U.S., even if American identity is not readily conferred to them [15, 80]. Shading reflects strength of association with Asian, Black, or Latina/o.

### Figure 3: BLIP American Classifications by Race or Ethnicity

Visual Question Answering As shown in Figure 3, posing the question "Is this person an American?" to the BLIP visual question answering model yields unambiguous results: 96.7% of White individuals are evaluated to be American, while only 2.8% of Asian individuals are evaluated to be American. 68.5% of Black individuals and 61.1% of Latina/o individuals are identified as American.

American Trait SC-EAT Table 3 indicates that, across all three visual semantic embedding spaces, White individuals are differentially associated with patriotism when compared with Asian, Black, or Latina/o individuals, with statistically significant effect sizes ($d$) ranging between .28 and 1.23. Only one effect size is not statistically significant, for the White vs. Latina/o comparison in SLIP. Table 4 shows that, in BLIP and CLIP, Asian, Black, and Latina/o are strongly associated with egalitarianism when compared to White individuals, represented with the phrase "a photo of someone who treats people of all races and backgrounds equally." This finding does not hold for SLIP. Table 5 shows that, across all three models, Asian or Latina/o individuals are strongly associated with "a photo of someone who is an immigrant to America" when compared with White individuals. The effect holds for Black individuals vs. White individuals in BLIP, but not in CLIP or SLIP. The findings reflect those of Devos and Banaji [15], who found that White individuals are perceived to be more patriotic; that Asian individuals are perceived to have been born outside of the U.S.; and that White individuals are not as likely to be perceived as treating people of all races and backgrounds equally, despite the association of American identity with both egalitarianism and being White.

Downstream Language-and-Image Tasks Results indicate that biases associating White with American in language-and-image AI extend beyond visual semantic embedding spaces, and introduce bias into downstream tasks including visual question answering, image captioning, and synthetic image generation.
Asian and Latina women are described as "oriental" in automatically generated captions, reflecting that the model fails to integrate Asian individuals into American identity. BLIP responds with a U.S. state for all White individuals.

toward Asian individuals is pronounced enough that the model does not associate the word "state" with an American state, which is the case for more than 97% of the other images in the CFD. Similarly, the model answers that 3.1% of Black individuals live in South Africa. New York seems to be a default state, and over-association with certain states occurs again in this experiment: BLIP answers South Carolina for 46.7% of Black individuals, and California for 33.9% of Asian and 63.9% of Latina/o individuals.

**Image Captioning** At every level of top-$p$ between .5 and .9 (the model has access to a greater number of sentence continuations as top-$p$ increases), the race of Asian individuals is the most commonly remarked upon by BLIP, with race noted between 12.84% and 36.70% of the time. The race of Black individuals is the second most remarked upon, between 2.03% of the time and 18.27% of the time. Race is never remarked upon for images of White individuals, and is only noted for Latina/o individuals when the model identifies an image of an individual who self-identifies as Latina/o to be Asian or Black. While Black individuals are described as "African American" 92.9% of the time when race is described, Asian individuals are always described as "Asian," and never as "Asian American."

While less common, the captions automatically generated by BLIP also reflect specific racial and gender associations and biases. In several troubling instances directly relevant to this research, Asian and Latina women are described as "oriental" in automatically generated captions, an offensive word directly highlighting the perceived foreignness of an individual [55]. Latina/o individuals are sometimes described as "Asian," potentially indicating that the model forms a broad and not always distinguishable representation of racial outgroups. Moreover, generated captions sometimes describe Black individuals as having an afro or dreadlocks, when these individuals do not in fact have these hairstyles. Despite all of the photographs in the CFD being headshots, the generated captions sometimes remark on the chest size of women, and may describe whether they are wearing makeup, or if their skin is oily. Despite all subjects wearing an identical grey t-shirt visible only at the top of the shoulders, men are sometimes described as wearing a black tie, reflecting the association of men with professional environments.

**Synthetic Image Generation** In all cases, the CLIP-guided VQGAN increases the mean pixel brightness of the skin, indicating that skin tone has been lightened in response to the text "an American person." For Black individuals, there is a 35% increase in mean pixel brightness over the first 80 training iterations, from 134.84 at the beginning of training to 181.89. However, the lightening of skin tone occurs even when the initialization image reflects an individual who self-identifies as White. Mean pixel brightness increases by 13% for Asian individuals, by 20.4% for Latina/o individuals, and by 15.7% for White individuals at the highest brightness. Moreover, the data suggest that the generative model does not immediately locate the direction which lightens skin tone. In the first ten training iterations, the brightness of skin tone actually decreases for Asian and Black individuals. However, by the twentieth training step, the model finds that skin tone is the most impactful direction for matching the output image to the text input. Lightening of skin color continues until the eightieth step, when the model begins to add features such as American flags and bald eagles to the image. While the brightness of skin tone decreases after the one hundredth training step, qualitative inspection reveals that the model is not making skin tone darker; rather, the cropped region of the forehead tends to become obscured by an American flag, or by long blonde hair. Images later in the series appear to try to increase the realism of the depicted person by adding lines and shading to the forehead.

6 **DISCUSSION**

Among the most cherished explicitly stated values of Americans are a commitment to the equality of all people without regard to
race or ethnicity, and a belief that all people who make a home in America have an equal claim to American identity [23, 58]. Yet expressed beliefs often do not reflect behavior, and experimental psychologists have found implicit attitudes excluding people who are not White from the category of American [15]. This research demonstrates that these biases have been learned by AI, and that the implicit biases present in multimodal embedding spaces impact the biases which manifest in downstream language-and-image tasks.

The correlation between state-level IAT scores and the number of images of Black individuals returned by state in CLIP suggests that a connection exists between bias in human subjects and the threat to White prototypicality in a region, in keeping with the research of Danbold and Huo [12], who find that fear of losing prototypical status predicts lower support for diversity among Americans. The results suggest that, where a state is more associated on the national or international level with Black individuals than with White individuals, White racial biases are stronger.

Tests assessing biases using the BLIP visual question answering head also suggest that American identity is not conferred upon any individual living in the U.S. While the model responds that only 61.1% of Latina/o individuals are Americans, it also responds that 63.9% of Latina/o individuals live in the state of California, and that 26.9% of Latina/o individuals live in New York. Unlike for images of Asian individuals, for whom BLIP does not make the inference that an American state is being requested, BLIP is able to assign residence in an American state to Latina/o individuals - but not American identity. A similar result is observed for Black individuals, as the model responds that 68.5% of Black individuals are American, but also responds that 46.7% of Black individuals live in the state of South Carolina, and 46.7% of Black individuals live in the state of New York. This incongruence suggests that to be American connotes something more in multimodal AI than living in the U.S. Moreover, that the model infers in most cases that the word "state" refers to an American territory, rather than to a nation-state, reflects an American-centric bias in the massive webscraped corpora used to train language-and-image AI.

Finally, a text-generation task renders the race of Asian, Black, and Latina/o individuals more visible than the race of White individuals, while an image-generation task renders visible only White individuals, even when provided with initialization images of Asian, Black, and Latina/o individuals. The BLIP image captioning model is more likely to draw attention to the race of Black individuals and especially Asian individuals, remarking upon the race of Asian individuals 36.7% of the time when top-$p$ is set to .7. On the other hand, a CLIP-guided VQGAN lightens skin tone by at least 13% based on pixel brightness for all races and ethnicities when provided with the text guidance "an American person," and by 35% for initialization images of Black individuals. These results both derive from the default association of American identity with being White. Where only text referring to the category American is given, the most associated race is White. Where a model detects difference from the default, the deviation is described in racialized terms.

This research suggests the potential impact of AI on humans, as racial prejudice related to exclusion from American identity has been shown to cause depression and low self-esteem [2, 28]. Multimodal models similar to those examined in this research have been proposed as the future of ubiquitous internet applications such as Search [40], and image generators capable of photorealistic output are expected to be widely usable by practitioners in the near future [41]. Such systems are trained on internet-scale web searches, and are thus likely to encode societal biases consistent with those identified in this research. Unchecked, exposure to AI-generated content reflecting the prototypical association of American identity with being White may amplify a bias observed in humans.

Limitations and Future Work This research does not conduct a thorough evaluation of the training data for the models examined, and the training data for CLIP is not publicly available [47]. Previous work analyzing dataset composition has yielded insight into the causes of machine bias [3, 17, 74], and future work might systematically explore the contents of those language-and-image AI training datasets which are publicly available. Moreover, this research examines three English-language models, because it is primarily concerned with whether an American bias is encoded into language-and-image AI. Should multilingual models comparable to CLIP be made available for research, they might be studied for the presence of similar ethnocentric biases. Additionally, language-and-image AI relies on language models to encode representations of text, which are inevitably sensitive to context. While this research defines prompts based on a principled method suggested by the designers of the models studied, it is unavoidable that changing the prompt will produce different embedding association results in some circumstances. Finally, future work might explore how other identity-related prototypicality associations manifest in AI.

7 CONCLUSION

In accordance with findings from experimental psychology [15], the present research reveals systematic biases in CLIP, SLIP, and BLIP associating American identity with being White. From biases associating White individuals with in-group words in visual semantic embedding spaces to the unambiguous exclusion of Asian, Latina/o, and Black individuals from American identity in visual question answering tasks, the results indicate that language-and-image AI learns that the prototypical American is White.

ACKNOWLEDGEMENTS

This material is based on research partially 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 authors and do not necessarily reflect those of NIST.
A. APPENDIX

A.1 Foundational Research in Language-and-Image AI

The ideas central to the design of modern language-and-image AI, including the use of natural language supervision to learn visual semantic features, were first introduced by Socher et al. [61], and advanced by Frome et al. [20] in the design of the DeViSE deep learning visual semantic embedding model. While foundational for visual semantics, these models did not approach the performance of CLIP, which improved the zero-shot state-of-the-art on the ImageNet benchmark [13] from 11.5% [34] to 76.2% [47]. The contrastive learning objective was introduced by Tian et al. [64], and CLIP builds most directly on the ConVIRT medical image classifier of Zhang et al. [79], and was designed concurrently with the zero-shot ALIGN language-image model of Jia et al. [29]. Most recently, Tiwary [65] introduce Turing-NLG, a multilingual visual semantic model that outperforms CLIP on zero-shot image classification. Recent research suggests that visual semantic pretraining may have benefits for language representations, independent of image representations, as the word and sentence embeddings formed by CLIP have been shown to be highly semantic [75], setting or matching state of the art on the RGE5 [54] and ValNorm [66] intrinsic evaluations.

A.2 WEAT Formulae

As described by Caliskan et al. [7], the formula for the WEAT is given by:

$$\text{WEAT}(x, A, B) = \text{mean}_{x \in X} (\text{mean}_{y \in Y} (X, A, B)) - \text{mean}_{y \in Y} (X, A, B)$$

where the association for a word $w$ is:

$$\text{mean}_{w \in W}(\text{mean}_{x \in X} (\text{mean}_{y \in Y} (X, A, B)))$$

The SC-WEAT measures the directional angular similarity between two groups of attribute words with a single target word. The formula for the SC-WEAT is given by:

$$\text{SC-WEAT}(x, A, B) = \text{mean}_{w \in W}(\text{mean}_{x \in X} (\text{mean}_{y \in Y} (X, A, B)))$$

The models of Jia et al. [29] and Tiwary [65] are not publicly available to researchers.
Figure 8: Images generated from the "an American person" text prompt with no initialization image, using only text guidance. Commensurate with the research of Goh et al. [21], the model often generates a visual depiction of the input text.