

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|--|
| n/a | Confirmed |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of all covariates tested |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

- | | |
|-----------------|---|
| Data collection | Our team developed custom Python code (Python version 3.12.0) for collecting the observational data from Google and Wikipedia reflected in panels A, B, and C of figure 1. The remaining data reflected in panels D-J of figure 1 were already collected and published by other teams, and we detail and cite these sources directly in the main text (many of these datasets are established training sets for computer vision models). Except for the word2vec model we trained on a recent sample of news we collected using Crawl Feeds (see figure S14), all other language models examined are either publicly available via Python packages (e.g., gensim) or via publicly available APIs as in the case of GPT models. The experimental data reflected in figure 3 were collected using a survey instrument developed in Qualtrics and a panel of participants from Prolific. |
| Data analysis | All data analyses were conducted using custom code written by our team in either R or Python (Python version 3.12.0; R version 4.4.2). All data and code for replicating the analyses in our paper are publicly available at the following github: https://github.com/drguilbe/distortion_age_gender_online . Our statistical analyses do not control for multiple comparisons since all tests were theoretically motivated and, when experimental, preregistered -- and none of our analyses involved an agnostic permutation over a set of pairwise comparisons. |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All data and code associated with this study can be publicly downloaded at: https://github.com/drguilbe/distortion_age_gender_online

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

In this study, we examine "gender" as a socially constructed category specifying a manner of identification and not sex as a biologically determined phenotypic category. Our measurements of gender - either in images, videos, or text - are in no way intended or framed to reify gender as biologically determined or as an objectively detectable and static attribute of specific individuals or of people in general. Instead, we study gender as a perception -- i.e., as (an often biased) judgment of people or of social categories in general (e.g., occupations). Accordingly, we measure gender in a number of ways, none of which constitute a fully objective ground truth, especially when understanding gender as both a fluid mode of self-identification and as a context-dependent perception by others. In the context of images, we measure gender using (i) classification judgments aggregated across human coders, (ii) machine learning classifications, and (iii) self-identified gender ascriptions based on publicly available biographical profiles. In the case of language models, we study gender at the level of social categories (e.g., occupations) by examining the extent to which a given category is associated with men or women in these language models' high dimensional embedding space. We adopt a binary partitioning of gender (female and male) when extracting gender associations from language models not because we believe or intend to convey an essentialist view of gender as intrinsically binary. We maintain that gender is highly fluid and exists in a complex, fluid, multidimensional continuum. However, in order to compare our image data against established methods for measuring gender associations in language models (specifically, the "geometry of culture" method), we adopt the binary partitioning of gender in embedding space that this prior work uses and validates via a series of robustness tests.

Reporting on race, ethnicity, or other socially relevant groupings

In this study, we do not directly examine socially constructed categorizations relating to race or ethnicity. In our resume audit study of ChatGPT, we adopt the methods of recent work which provided a set of names for men and women that were normalized by familiarity and chosen to be representative across the following four ethnic categories (according to their categorizations): (Black or African American, Hispanic or Latinx, Asian, and White. See the original paper that provided this set of names here: <https://arxiv.org/abs/2405.04412>

Population characteristics

For this experiment, we recruited a nationally representative sample of participants (n = 500) from the popular crowdsourcing platform Prolific, which provides a panel of high-quality human participants for online research. Our sample size was selected to emulate the sample size of a recent experiment with a highly similar design, which effectively measured statistically powered outcomes (see Guilbeault et al., 2024 in Nature). 459 participants completed the task, exhibiting an attrition rate of 9.2%. We only examine data from participants who completed the experiment. To recruit a nationally representative sample, we used Prolific's pre-screening functionality designed to provide a nationally representative sample of the USA along the dimensions of sex, age, and ethnicity. Participants were invited to partake in the study only if they were based in the USA, fluent English speakers and aged more than 18 years. A total of 52% of participants were female (no participants identified as non-binary). The average age of participants was 45.2 (45.9 for women, 44.6 for men). All participants provided informed consent before participating. This experiment was run on November 10th, 2023.

Recruitment

For the experimental component, our sampling strategy was a random sample from Prolific's nationally representative panel (n = 500; see "Population Characteristics").

Ethics oversight

This study was approved by the Ethics Review Board IRB at the University of California, Berkeley, where this study was conducted.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study has three main components. The first is an observational analysis of the gender and age associations of social categories in images, videos, and textual data from popular online platforms, including Google, Wikipedia, Flickr, IMDb (Internet Movie Database), and Youtube. The second component is an online experiment in which human participants were tasked with using Google Images to search for images of occupations. This experiment tested the effects of exposure to online images of occupations on people's beliefs about the age and hireability of men and women in these occupations. The third component of this study is a resume audit of ChatGPT. For this audit, we prompted ChatGPT to generate resumes across a number of occupations (yielding over 40k resumes), while varying whether the resume was generated for a male or female applicant (based on the name of the target applicant). We then measured how varying the name of the applicant impacted the age and level of experience that ChatGPT assigned to this applicant, as well as ChatGPT's overall score for the quality of the resume it generated.
Research sample	<p>The research sample for the experimental component of this study consists of a national representative sample of the U.S. as curated by the crowdsourcing platform Prolific. The details on Prolific's U.S. nationally representative sample are provided by Prolific at the following link: https://researcher-help.prolific.com/en/article/95c345</p> <p>To create a representative U.S. sample, Prolific takes the intended sample size and strategies it across three demographics: age, sex, and ethnicity. Prolific uses census data from the U.S. Census Bureau to divide the sample into subgroups with the same proportions as the national U.S. population. This means, for example, that a nationally representative sample contains the same proportion of 28-37 year old Asian women as the national population (to the extent possible). Using this representative sample is important for our experiment, which does not make any demographic-specific predictions around the effects of Google Image search on gender-age bias. In this way, based on these available resources, our findings are positioned to generalize across the aforementioned demographic characteristics.</p>
Sampling strategy	<p>For the experimental component of our study, our sampling strategy was a random sample from Prolific's nationally representative panel (n = 500, see "Research Sample"). In terms of the statistical method used to determine sample size, this sample size replicated the sample size from a recent publication by our team which presented statistically significant effects of Image search with a comparable sample size (see Guilbeault et al. 2024 in Nature). Otherwise, no power tests were used to select this sample size; we based our judgment on the prime facie validity of reproducing this recent sample size from Guilbeault et al. 2024.</p> <p>For the observational component of this study, our sample of Google images was collected through the following standardized procedure (this description is copied from the Guilbeault et al. 2024 Nature reporting summary, which first introduced this dataset). We started by using each of the social categories in Wordnet to automatically search and retrieve the top 100 images in Google corresponding to each social category in Google Images (Google provides roughly 100 images by default for its initial search results on a given search query). Each search was implemented from a fresh Google account with no prior history to avoid the uncontrolled effects of Google's recommendation algorithm, which customized search results based on browsing history. Searches were run by 10 distinct servers in New York City. All image data from Google was collected in August 2020.</p> <p>The sources of the language models are transparently described in the paper and appendix. All models were available via Python (e.g., <code>genism</code> package) or via public API in the case of GPT models. One of the models we examine in the appendix was trained on a recent sample of online news that our team collected. We compiled a dataset of 2,717,000 randomly sampled news articles published in English across various topics between January 2021 and August 2023. These articles were sourced from the following prominent online news sources: 1,000,000 articles from the BBC; 500,000 articles from the Huffington Post; 480,000 articles from CNBC; 400,000 articles from Bloomberg; 160,000 articles from Time Magazine; 150,000 12 articles from Techcrunch; and 27,000 articles from CNN. These datasets were purchased from the online web-scraping service, Crawl Feeds (https://crawlfeeds.com/).</p>
Data collection	The online observational data – text, images, and videos – all derived from publicly available repositories. The experimental data collected was implemented using a survey instrument designed in Qualtrics and deployed over Prolific.
Timing	All of our main image data from Google was collected in August 2020. The timing associated with the training and release of all language models is described in the manuscript and appendix; similarly for the timing associated with the already published image and video training sets that we examine. One of the models we examine in the appendix was trained on a recent sample of online news that our team collected. We compiled a dataset of 2,717,000 randomly sampled news articles published in English across various topics between January 2021 and August 2023. The experiment reported in figure 2 was conducted on November 10th, 2023. The resume audit study of ChatGPT was conducted between July and August of 2024.
Data exclusions	No data were excluded from the observational analyses. For our experimental data, we only examined data associated with participants who successfully completed the task (459/500).
Non-participation	No participants declined to participate in this task based on our records.
Randomization	In the experiment, participants were evenly randomized into either the control or the image condition.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern
<input checked="" type="checkbox"/>	<input type="checkbox"/> Plants

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Plants

Seed stocks	<input type="text" value="NA"/>
Novel plant genotypes	<input type="text" value="NA"/>
Authentication	<input type="text" value="NA"/>