
Supplementary information

**Age and gender distortion in online media
and large language models**

In the format provided by the
authors and unedited

Supplementary Appendix for:

Age and Gender Distortion in Online Media and Artificial Intelligence

Douglas Guilbeault^{1*}, Solene Delecourt², Bhargav Srinivasa Desikan³

¹Graduate School of Business, Stanford University, Stanford, CA, 94117

²Haas School of Business, University of California Berkeley, Berkeley, CA, 94720

³Autonomy Institute, London, United Kingdom, E1 5QJ

***Correspondence:** dguilb@stanford.edu

Project github: https://github.com/drguilbe/distortion_age_gender_online/

Table of Contents

A. 1. Extended Census Analyses.....	pg. 1
B. 1. 1 Robustness to Search Location.....	pg. 5
B. 1. 2 Robustness to Statistical Measure of Age in Human Classifications.....	pg. 6
B. 1. 3 Robustness to the Demographics of Human Coders.....	pg. 8
B. 1. 4 Evaluating the Accuracy of Human Coder Judgments.....	pg. 10
B. 1. 5 Evaluating Inter-coder Agreement in Gender and Age Classification.....	pg. 13
B. 1. 6 Sensitivity Analyses Demonstrating Robustness to Potential Coder Biases.....	pg. 14
B. 1. 7 Robustness to Controlling for Linguistic Features of Social Categories.....	pg. 16
B. 1. 8 Robustness to Image Features.....	pg. 17
B. 1. 9 Robustness to Non-cropped Image Classification.....	pg. 18
B. 1. 10 Validation of the OpenCV Image Cropping Algorithm.....	pg. 20
B. 1. 11 Sources of Images in Google Image Dataset.....	pg. 21
B. 1. 12 Benchmarking Image Data Against the U.S. Census.....	pg. 22
B. 1. 13 Robustness of Status Associations in Google Images.....	pg. 25
C. 1. 1 Robustness of the Geometry of Culture Method at the Sentence-level.....	pg. 30
C. 1. 2 Robustness of GPT2-Large Results Controlling for Linguistic Features of Categories.....	pg. 34
C. 1. 3 Replication across Nine Language Models.....	pg. 36
C. 1. 4 Robustness of the Age Dimension across Models.....	pg. 38
C. 1. 5 Benchmarking GPT2-Large Demographic Dimensions Against the Census.....	pg. 39
D. 1. 1 Robustness to Controlling for Participants' Gender and Age.....	pg. 42
D. 1. 2 Robustness to Alternative Experimental Measures of Gender Association.....	pg. 46
D. 1. 3 Robustness to Likert Measure of Perceived Hireability.....	pg. 48
D. 1. 4 Benchmarking Participants' Age and Gender Judgments Against the U.S. Census Data.....	pg. 49
D. 1. 5 Full Summary of Pre-registered Hypotheses and Results.....	pg. 51
E. 1. 1 Coherence of ChatGPT's Responses.....	pg. 54
E. 1. 2 Comparing ChatGPT's Age Assignments for Occupations to U.S. Census Data.....	pg. 55
E. 1. 3 Robustness to Model Temperature.....	pg. 56

Supplementary Analyses

A. 1. Extended Census Analyses

Here, we provide more detailed statistical analyses of the U.S. Census data examined as part of setting the stage for our broader inquiry. We begin by visualizing the median age of women and men throughout American society (across all age groups, both employed and unemployed) in Fig. S1. The data shows that for every single year going back to 2009, American women in the U.S. population are older than men, and this age gap has remained strikingly stable. These analyses are based on the “gender_table1” file made available for each year at www.census.gov.

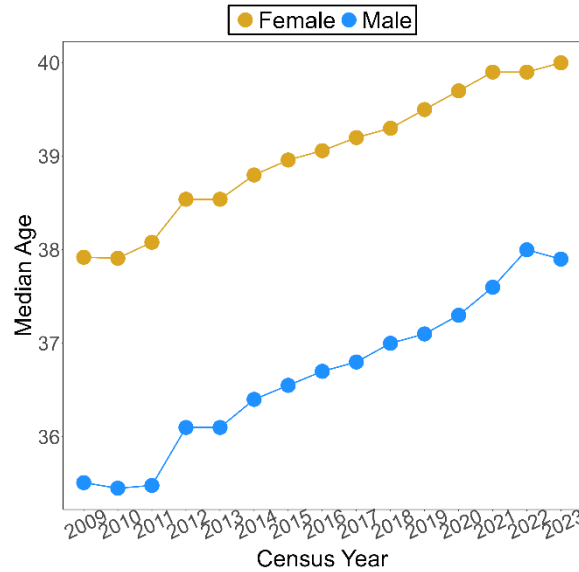


Figure S1. The median age of women and men (across all age groups, both employed and unemployed) in the U.S. census from 2009 to 2023.

Next, we present a statistical model demonstrating the robustness of the results presented in Extended Figure 1, which indicates the lack of correlation between the fraction of women in an occupation and the median age of people in this occupation in (A) 2012 and (B) 2023, according to the U.S. census. Table S1 displays an OLS regression that predicts the median age of an occupation as a function of the fraction of women employed in this occupation, while controlling for fixed effects by census year (from 2011 to 2023), and while clustering standard errors at the occupation level. This date range includes all census tables made publicly available with the associated statistical information. This specific analysis was enabled by matching all “cpsaat11b” tables (which indicate the median age associated with each occupation) with all “cpsaat11” tables (which indicate the fraction of women employed in each occupation) from www.bls.gov. The results indicate that there is no significant correlation between the fraction of women in an occupation and its associated median age ($\beta = -0.01$, $CI = [-0.02, 0.00]$, $p = 0.14$). In addition to yielding no statistically significant relationship among these variables of interest, the model also captures a vanishingly small amount of the variance as indicated by the adjusted R^2 of 0.002 (i.e., capturing less than 1% of the variance of median age).

Variables	Beta	95% CI ¹	p-value
% Women	-0.01	-0.02, 0.00	0.13707
Census Year			—
2011	—	—	—
2012	0.27	0.05, 0.47	0.01416
2013	0.48	0.25, 0.69	1.993e-05
2014	0.04	-0.21, 0.30	0.73157
2015	0.25	0.00, 0.49	0.04364
2016	0.33	0.07, 0.58	0.01212
2017	0.13	-0.14, 0.41	0.35817
2018	0.25	-0.04, 0.54	0.09204
2019	0.17	-0.13, 0.46	0.28094
2020	0.51	-0.00, 1.03	0.05445
2021	0.23	-0.31, 0.78	0.41038
2022	0.11	-0.44, 0.65	0.70485
2023	-0.10	-0.66, 0.47	0.73724
Statistic	1.80		—
R ²	0.005		—
Adjusted R ²	0.002		—
No. Obs.	4,700		—
df	13		—
AIC	29,224		—
Residual df	4,686		—
Sigma	5.41		—

¹CI = Confidence Interval

Table S1. An OLS regression predicting the median age associated with an occupation in the U.S. census as a function of the fraction of women in the occupation, while controlling for census year fixed effects (from 2011 to 2023, the date range for which the properly formatted data is available), and while clustering standard errors at the occupation level ($N = 725$ occupations).

Finally, we provide the full details on the statistical analyses of the census data indicating minimal age differences between men and women across the entire workforce. This analysis is based on the “gender_table7” data from 2013 to 2023, as well as the alternatively named “gender_table13” for 2012 and “gender_table16” for 2009-2011. All tables are publicly available and were downloaded from www.census.gov. We restrict our analyses to this date range because our interest is primarily in documenting already established trends between gender and age in the workplace roughly over the most recent decade. Within this time window, the report trends are remarkably stable. The tables examined from the U.S. census bureau of labor statistics separately indicate the fraction of the male workforce and the fraction of the female workforce that are distributed across the following age bins:

“16-17”, “18-24”, “25-29”, “30-34”, “35-39”, “40-44”, “45-49”, “50-54”, “55-59”, “60-64”, and “65+”. These tables present fully aggregated results at the level of the entire workforce with no breakdown by industry or occupation.

Variables	Beta	95% CI ¹	p-value
Gender			—
Female	—	—	—
Male	0.00	-0.01, 0.00	1.897272e-01
Age Group			—
16-17	—	—	—
18-24	0.11	0.10, 0.11	1.591314e-149
25-29	0.10	0.09, 0.10	2.760839e-141
30-34	0.10	0.09, 0.10	2.695797e-138
35-39	0.09	0.09, 0.10	7.843504e-135
40-44	0.09	0.09, 0.10	8.813223e-137
45-49	0.10	0.09, 0.10	3.537072e-141
50-54	0.10	0.09, 0.10	1.247825e-141
55-59	0.09	0.08, 0.09	2.547033e-127
60-64	0.06	0.05, 0.06	6.311533e-87
65+	0.04	0.04, 0.05	3.147040e-61
Census Year			
2009	—	—	—
2010	0.00	0.00, 0.00	1.000000e+00
2011	0.00	0.00, 0.00	1.000000e+00
2012	0.00	0.00, 0.00	1.000000e+00
2013	0.00	0.00, 0.00	1.000000e+00
2014	0.00	0.00, 0.00	1.000000e+00
2015	0.00	0.00, 0.00	1.000000e+00
2016	0.00	0.00, 0.00	1.000000e+00
2017	0.00	0.00, 0.00	1.000000e+00
2018	0.00	0.00, 0.00	1.000000e+00
2019	0.00	0.00, 0.00	1.000000e+00
2020	0.00	0.00, 0.00	1.000000e+00
2021	0.00	0.00, 0.00	1.000000e+00
2022	0.00	0.00, 0.00	1.000000e+00
2023	0.00	0.00, 0.00	1.000000e+00
Gender x Age			

Variables	Beta	95% CI¹	p-value
Male x 18-24	-0.01	-0.01, 0.00	2.761436e-02
Male x 25-29	0.01	0.00, 0.01	7.904323e-02
Male x 30-34	0.01	0.00, 0.01	6.078730e-03
Male x 35-39	0.01	0.00, 0.01	3.988876e-03
Male x 40-44	0.00	0.00, 0.01	1.347623e-01
Male x 45-49	0.00	0.00, 0.01	6.995826e-01
Male x 50-54	0.00	-0.01, 0.01	8.556361e-01
Male x 55-59	0.00	-0.01, 0.00	7.830493e-01
Male x 60-64	0.00	0.00, 0.01	7.436885e-01
Male x 65+	0.01	0.00, 0.02	1.270018e-03
<i>Statistic</i>	288		
<i>R²</i>	0.972		
<i>Adjusted R²</i>	0.968		
<i>No. Obs.</i>	330		
<i>df</i>	35		
<i>AIC</i>	-2,454		
<i>Residual df</i>	294		
<i>Sigma</i>	0.006		

¹CI = Confidence Interval

Table S2. An OLS regression predicting the fraction of people employed in the workforce as a function of an interaction between gender and age group of people, while also including fixed effects by census year (2009 to 2023).

Table S2 presents an OLS regression that predicts the fraction of people employed in the workforce as a function of an interaction between gender and age group, while also including fixed effects by census year. This model effectively tests whether men are more likely to belong to a particular age group than women, controlling for census year. The relevant correlations of interest in Table S2 are highlighted in bold. The model shows that while employed men are less likely to belong to the age range of 18-24 compared to women ($\beta = -0.01$, CI = [-0.01, 0.00], $p = 0.03$), men are consistently more likely to belong to the age groups of 25-29 ($\beta = 0.01$, CI = [0.00, 0.01], $p = 0.03$), 30-34 ($\beta = 0.01$, CI = [0.00, 0.01], $p = 0.006$), and 35-39 ($\beta = 0.01$, CI = [0.00, 0.01], $p = 0.003$). Yet, importantly, there is no significant difference in the likelihood of men and women belonging to the age groups 40-44, 45-49, 50-54, 55-59, 60-64 (if anything, for ages 50-60, the direction of the coefficient is negative in the direction of men being less likely). Figure S2 visually plots the partial effect of the interaction of gender and age on the fraction employed, while maintaining the controls in Table S2.

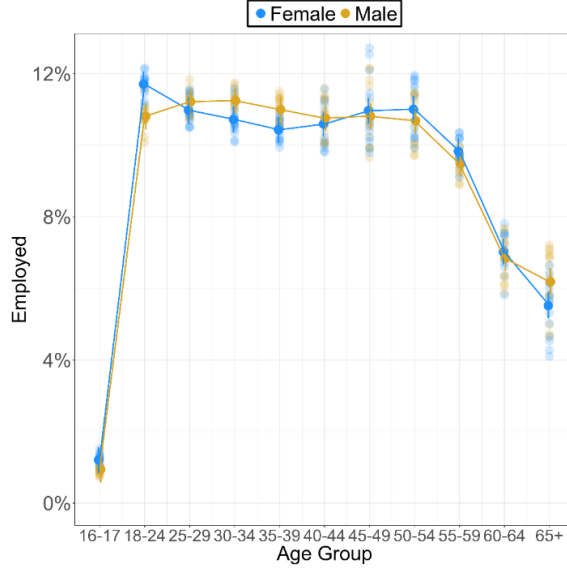


Figure S2. Partial effect plot predicting the fraction of people in the workforce according to the U.S. census as a function of the interaction between age group and gender, while including year fixed effects for each census year from 2009 to 2023. See Table S2 for full statistical analysis. Bold data points show the average fraction percentage of employed people within each age group averaged across years. Error bars show 95% confidence intervals for the average across years. Faded data points show the fraction of employed people within each group for each year separately.

It is outside the scope of this study to provide an account of what is driving these patterns in the U.S. workforce. Our aim in examining this publicly available U.S. census data is to highlight the ways in which the available ground truth data fall short of providing clear evidence of systemic age gaps between men and women in the workforce and beyond. We provide direct comparisons of age-gender associations between our Google image data on occupations and census data on industry-level averages in Extended Data Figure 2 and 3.

B. 1. Image Analyses

B. 1. 1 Robustness to Search Location

Here, we examine the robustness of our results when altering the IP from which the Google Images are searched and collected. We run this analysis for a random sample of 300 social categories from the full set of 3,435 Wordnet social categories examined in our main paper. The same categories were searched across all IPs. Searching from separate locations led to distinct images in Google’s search results: on average, less than 30% of the images associated with the same social category repeated across geolocated searches.

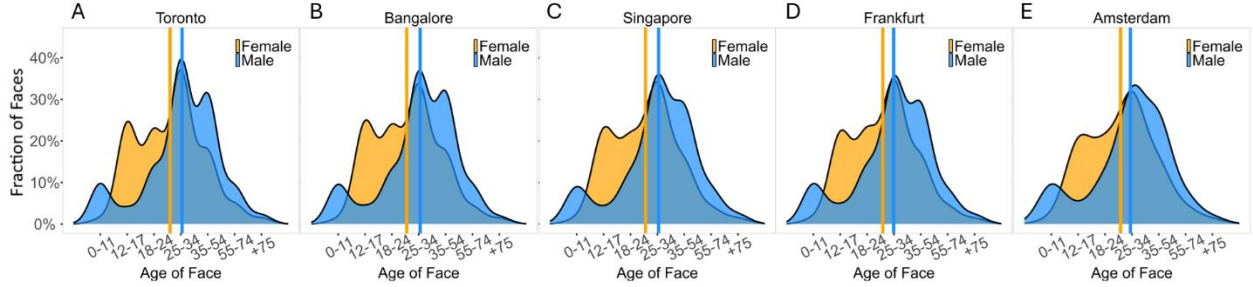


Figure S3. Human judgments of age and gender are shown for the same 300 social categories in Google Images collected while searching from an IP address based in (A) Toronto ($n = 9,902$ images), (B) Bangalore ($n = 8,058$ images), (C) Singapore ($n = 8,334$ images), (D) Frankfurt ($n = 9,501$ images), and (E) Amsterdam ($n = 8,644$ images). Gold (blue) lines indicate the average age for female (male) faces according to each dataset.

Yet, figure S3 shows that despite this variation, the correlation between women and younger ages held across all IPs. Women were identified as significantly younger than men in Google images collected by searches run in Toronto (Figure S1A, $t = -19.48$, $CI = [-0.49, -0.40]$, $p = 2.2 \times 10^{-16}$, $n = 9,902$ images), Bangalore (Figure S1B, $t = -19.37$, $CI = [-0.54, -0.44]$, $p = 2.2 \times 10^{-16}$, $n = 8,058$ images), Singapore (Figure S1C, $t = -19.43$, $CI = [-0.53, -0.43]$, $p = 2.2 \times 10^{-16}$, $n = 8,334$ images), Frankfurt (Figure S1D, $t = -17.11$, $CI = [-0.45, -0.36]$, $p = 2.2 \times 10^{-16}$), and Amsterdam (Figure S1E, $t = -14.54$, $CI = [-0.42, -0.32]$, $p = 2.2 \times 10^{-16}$, $n = 8,644$ images), (Student's t-test, two-tailed). These findings indicate that our Google results are not an artifact of geolocation and instead reveal systemic bias in the age-related representation of women and men in online images from across the world.

B. 1. 2 Robustness to Statistical Measure of Age in Human Classifications

Here, we show that our main results using human coders are robust to a range of statistical approaches. In the main text, our results examining human-coded images from Google and Wikipedia examine the average age judgment across three coders for each image. First, we show that these results replicate strongly if we examine the modal, rather than the average, age judgments across three coders (note, this approach has the limitation that in 12% of cases, faces were associated with three unique age bins, one per coder; in this case, an age bin is randomly selected).

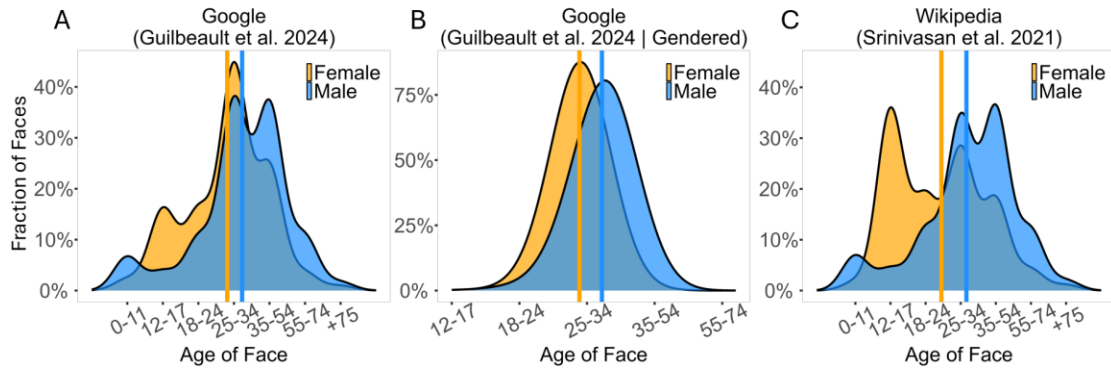


Figure S4. Women are represented as significantly younger than men in a large sample of images from Google and Wikipedia. The modal age (aggregated across three coders) of either female or male faces according to (a) the top 100 Google images associated with 3,434 social

categories ($n = 161,484$ images), (B) the top 100 Google images retrieved using gendered searches (e.g., by searching “female athlete” or “male athlete”) shown for all non-gendered categories in WordNet (2,960 categories; $n = 495,551$ images), (C) 1,251 categories in Wikipedia (from the Srinivasan et al. 2021 Wikipedia-based Image Text dataset; $n = 14,709$ images). Gold (blue) lines indicate the average age for female (male) faces according to each dataset.

We analyze a body of 657,035 images from the Google search engine associated with 3,489 social categories, where the age and gender classification of each face are represented using the modal judgment across three unique coders (in contrast to the main text where the average age judgment across coders is used). We find that women in Google images are coded as significantly younger than men both for non-gendered searches (e.g. searching “doctor” or “banker”; $t = -69.5$, $p = 2.2 \times 10^{-16}$, $n = 3,434$ categories; Fig. S4A) and gendered searches (e.g. searching both “female doctor” and “male doctor”; $t = -38.9$, $p = 2.2 \times 10^{-16}$, $n = 2,960$ categories; Fig. S4B). We replicated this methodology to analyze Wikipedia images. Again, we find that women in Wikipedia images are coded as significantly younger than men ($t = -36.72$, $p = 2.2 \times 10^{-16}$, $n = 1,251$ categories; Fig. S4C).

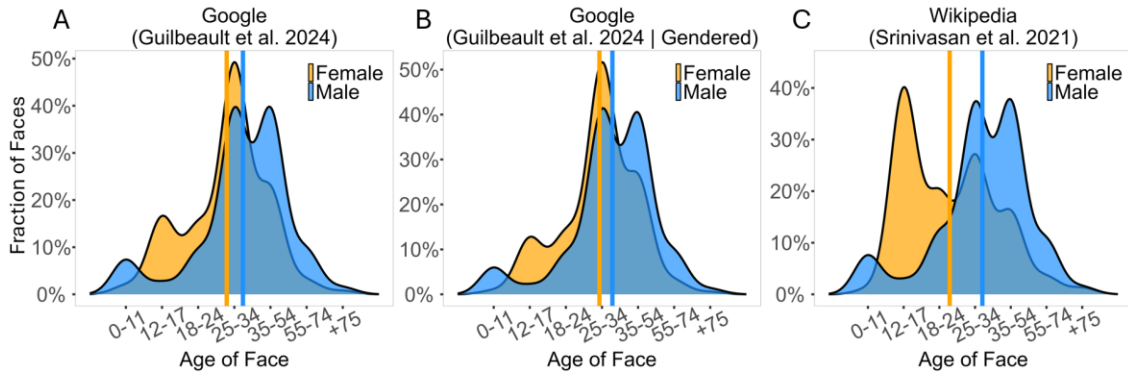


Figure S5. Women are represented as significantly younger than men in a large sample of images from Google and Wikipedia for which consensus on perceived age and gender was reached among three independent coders. The consensus age classification of either female or male faces according to (a) the top 100 Google images associated with 3,378 social categories ($n = 62,747$ images), (B) the top 100 Google images retrieved using gendered searches (e.g., by searching “female athlete” or “male athlete”) shown for all 2,959 non-gendered categories in WordNet ($n = 197,550$ images), (C) 1,068 categories in Wikipedia (from the Srinivasan et al. 2021 Wikipedia-based Image Text dataset; $n = 6,896$ images). Gold (blue) lines indicate the average age for female (male) faces according to each dataset.

We also test whether our results hold when only looking at images associated with perfect consensus in coders’ gender and age judgments. Full consensus on gender and age associations was reached for 64% of all images. Looking only at consensus images, we find that women in Google images continue to be coded as significantly younger than men both for non-gendered (e.g. searching “doctor” or “banker”; $t = -45.62$, $p = 2.2 \times 10^{-16}$, $n = 3,362$ categories, 57,021 images; Fig. S5A) and gendered searches (e.g. searching both “female doctor” and “male doctor”; $t = -70.81$, $p = 2.2 \times 10^{-16}$, $n = 2,957$ categories, 179,621 images; Fig. S5B). With the same technique, we similarly find that women in Wikipedia images continue to be coded as significantly younger than men ($t = -27.71$, $p = 2.2 \times 10^{-16}$, $n = 1,051$ categories, 6,553 images; Fig. S5C), (Student t-test, two-tailed).

B. 1. 3 Robustness to the Demographics of Human Coders

Here, we test whether the demographic composition of our annotator panel led to biases in how they classified the faces in our dataset. We report the results of an ordinary least squares regression that predicts the age classification of the face retrieved by Google Image search as a function of the coded gender of the face, with fixed effects by social category and the demographic features of the Mturk workers who coded the faces (including their gender, race, and age). Table S3 presents the results of this model. Table S3 indicates that our main result – whereby male faces are associated with significantly higher age ratings – continues to hold when controlling for coders’ demographics ($\beta[\text{Male}] = 0.42$, $\text{CI} = [0.41, 0.42]$, $p = 2.2 \times 10^{-16}$). The fixed effects associated with the demographic traits of Mturk workers were unrelated to the expected gender classification of faces, contributing to a marginal R^2 of only .019, while the model as a whole achieves a conditional R^2 of 0.186. This robustness analysis indicates that our results are unlikely to be driven by demographic-related biases in the gender classifications provided by our annotator pool.

Variables	Beta	95% CI ¹	p-value
Gender of Image			
Female	—	—	—
Male	0.42	0.41, 0.42	0.000000e+00
Data Source			
Google	—	—	—
Wikipedia	-0.15	-0.17, -0.14	7.965515e-106
Coder Age			
+75	—	—	—
18-24	0.08	0.04, 0.11	6.417344e-05
25-34	0.18	0.14, 0.21	1.019322e-20
35-54	0.15	0.12, 0.19	8.116869e-16
55-74	0.35	0.31, 0.39	1.004404e-61
Coder Gender			
Female	—	—	—
Male	0.07	0.07, 0.08	4.366002e-271
Non-binary	-0.37	-0.38, -0.36	0.000000e+00
Rather not say	0.16	0.14, 0.18	5.840102e-42
Coder Race			
African-American	—	—	—
Asian	0.26	0.26, 0.27	0.000000e+00
Caucasian	0.18	0.18, 0.19	0.000000e+00
Hispanic	0.08	0.07, 0.08	8.730075e-62
Native American	0.31	0.30, 0.32	0.000000e+00
Native Hawaiian	0.41	0.39, 0.44	3.212931e-226
Other	0.26	0.23, 0.30	9.102281e-49
Rather not say	0.34	0.31, 0.36	6.303894e-153
Two or more	0.32	0.29, 0.35	2.064554e-98
Two or more races	0.51	0.47, 0.55	3.086192e-120

Variables	Beta	95% CI ¹	p-value
Coder Income			
100k-150k	—	—	—
10k-50k	0.03	0.02, 0.03	1.385790e-14
150k-250k	-0.03	-0.04, -0.02	1.864658e-07
50k-75k	-0.05	-0.06, -0.04	4.621468e-46
75k-100k	-0.09	-0.10, -0.09	1.063985e-132
Less than 10k	0.02	0.01, 0.03	2.777032e-07
Over 250k	0.22	0.20, 0.24	2.905985e-85
Rather not say	-0.29	-0.34, -0.24	3.734708e-31
Coder Education			
Below High School	—	—	—
Doctorate (PHD)	0.26	0.24, 0.29	7.853577e-106
High School	0.09	0.08, 0.10	5.741608e-46
Master's	0.37	0.36, 0.38	0.000000e+00
Rather not say	0.03	-0.12, 0.18	6.947136e-01
Technical/Community College	0.22	0.20, 0.23	1.482055e-267
Undergraduate	0.29	0.28, 0.31	0.000000e+00
Coder Political Ideology			
Conservative	—	—	—
Conservative, Other	-1.3	-1.4, -1.3	2.086599e-226
Independent	0.03	0.02, 0.03	1.132599e-27
Liberal	0.03	0.02, 0.03	2.721176e-28
Liberal, Conservative	0.18	-0.17, 0.53	3.080463e-01
Other	0.21	0.20, 0.23	4.877281e-162
Rather not say	0.09	0.07, 0.10	9.211610e-40
<i>Statistic</i>	2,810		
<i>R²</i>	0.068		
<i>Adjusted R²</i>	0.068		
<i>No. Obs.</i>	1,424,754		
<i>df</i>	37		
<i>AIC</i>	4,460,915		
<i>Residual df</i>	1,424,716		
<i>Sigma</i>	1.16		

¹CI = Confidence Interval

Table S3. An OLS regression predicting coders' age classification of faces, which are treated as an ordinal numeric variable coded in the following way: (1) 0-11, (2) 12-17, (3) 18-24, (4) 25-34, (5) 35-54, (6) 55-74, (7) 75+. This model examines the correlation between this outcome variable and coders' gender classification, while controlling for the coders' own demographics in terms of their age, gender, race, income, education level, and political ideology.

B. 1. 4 Evaluating the Accuracy of Human Coder Judgments

Our main analyses of images from Google and Wikipedia rely on crowdsourced human judgments. To evaluate the accuracy of the gender and age judgments of these human coders, we developed a separate validation task where a subset of coders classified faces for which the true age and gender of each face is known. For this purpose, we used the IMDb-Wiki image dataset¹, which maps the birthdate and gender of celebrities (according to their IMDb and Wikipedia profiles) to time-stamped photos depicting these celebrities in order to infer the age of the celebrity as captured at the time of each photo. A strength of this dataset is that it contains over half a million faces, with substantial representation across all age groups considered in our study. For example, the Chicago Face dataset², a commonly used repository of images for research, only contains faces between 18 and 40 years old. Previous datasets used by prior studies to evaluate the accuracy of human judgments did not contain any infants, children, adolescents, or elderly people³.

For this validation task, we selected 5 male and 5 female faces from the IMDb-Wiki dataset for each of the seven age bins used in our analyses, totaling 70 unique images. We then recruited a random sample of 215 human coders from our original coder sample to classify the age and gender of each of these faces. Each coder classified 35 images randomly sampled from the entire set of 70. The order of images in this classification task was randomized. To control for familiarity biases, we asked each coder to indicate whether they were familiar with the face being classified. Only 5% of responses indicated familiarity. There was no significant difference in the extent to which participants indicated familiarity as a function of the gender of the face ($p = 0.09$, Wilcoxon Rank Sum Test, two-tailed), nor as a function of the age group of the face ($p = 0.14$, $\chi^2 = 9.7$, Kruskal-Wallis H Test, two-tailed). Our main results control for participants' familiarity judgments, and all of our main results hold if we exclude participant judgments that indicated familiarity with the face being classified.

Variables	Beta	95% CI ¹	p-value
Gender Classification			
Female	—	—	—
Male	-0.01	-0.11, 0.10	9.168072e-01
True Age			
0-11	—	—	—
12--17	1.2	1.1, 1.4	2.866770e-47
18-24	2.3	2.1, 2.4	9.282551e-147
25-34	1.9	0.99, 2.8	4.774395e-05
35-54	3.8	3.6, 3.9	0.000000e+00
55-74	4.1	4.0, 4.3	0.000000e+00
75+	5.5	5.4, 5.7	0.000000e+00
Familiar			
No	—	—	—
Not sure	-0.06	-0.11, -0.02	3.575722e-03
Yes	0.00	-0.06, 0.05	9.246493e-01
Race Classification			
African-American	—	—	—

Variables	Beta	95% CI¹	p-value
Asian	-0.06	-0.18, 0.07	3.704713e-01
Caucasian	-0.14	-0.25, -0.03	1.341519e-02
Hispanic	-0.10	-0.21, 0.01	8.143184e-02
Native American	-0.13	-0.25, -0.01	3.349115e-02
Native Hawaiian	-0.05	-0.18, 0.08	4.521178e-01
Other/Unknown	-0.12	-0.24, 0.00	4.866411e-02
Two or more	-0.11	-0.23, 0.00	4.981539e-02
Participant Fixed Effects			Included
Image Fixed Effects			Included
<i>Statistic</i>	405		
<i>R²</i>	0.945		
<i>Adjusted R²</i>	0.943		
<i>No. Obs.</i>	6,830		
<i>df</i>	279		
<i>AIC</i>	8,857		
<i>Residual df</i>	6,550		
<i>Sigma</i>	0.453		

¹CI = Confidence Interval

Table S4. An OLS regression predicting participants' judgments of the age of each face in the validation task, while controlling for the ground-truth gender and age group of the face. This model also includes fixed effects for each participant and image.

Across all age groups, we find that coders were able to accurately identify the self-identified gender of the person depicted 97% of the time. Similarly, coders' age judgments fell within one age group of the true age of each photo 98% of the time, with 68.5% of coders correctly identifying the precise age group. We used an OLS model to evaluate whether men and women were assigned different ages, while including fixed effects for the participant and the ground-truth gender and age group of the face being classified, in addition to clustering standard errors at the image level (table S4). We find no significant difference in subjects' age judgments as a function of the gender of the face being classified ($\beta[\text{Male}] = -0.01$, $\text{CI} = [-0.11, 0.10]$, $p = 0.92$). These outcomes all hold while controlling for whether participants indicated familiarity with the face being classified, where indicating familiarity was uncorrelated with subjects' age judgments ($\beta[\text{Yes}] = 0.00$, $\text{CI} = [-0.06, 0.05]$, $p = 0.92$). Similarly, these results hold when controlling for participants' judgments of the race of the face depicted.

Variables	Beta	95% CI¹	p-value
True Age	0.91	0.90, 0.92	0.000000e+00
True Gender			

Variables	Beta	95% CI ¹	p-value
Female	—	—	—
Male	0.05	0.00, 0.11	6.999287e-02
Familiar			
No	—	—	—
Not sure	-0.11	-0.16, -0.06	2.707736e-05
Yes	-0.04	-0.10, 0.02	2.005578e-01
Race Classification			
African-American	—	—	—
Asian	0.06	-0.01, 0.14	8.929658e-02
Caucasian	0.20	0.14, 0.26	4.174801e-11
Hispanic	0.30	0.22, 0.37	4.056634e-15
Native American	0.23	0.15, 0.31	1.911424e-08
Native Hawaiian	0.26	0.14, 0.38	2.266271e-05
Other/Unknown	0.34	0.25, 0.43	2.358900e-13
Two or more	0.18	0.10, 0.26	1.922821e-05
Participant Fixed Effects			Included
True Age * True Gender [Male]	0.00	-0.01, 0.01	9.733004e-01
<i>Statistic</i>	359		
<i>R²</i>	0.920		
<i>Adjusted R²</i>	0.917		
<i>No. Obs.</i>	6,830		
<i>df</i>	211		
<i>AIC</i>	11,333		
<i>Residual df</i>	6,618		
<i>Sigma</i>	0.546		

¹CI = Confidence Interval

Table S5. An OLS regression predicting participants' perceptions of the age of each face in the validation task as a function of the true age and gender of the face, as well as the interaction of these predictors. Fixed effects are included for participant, the race of the target face, and whether participants were familiar with the face. In this model, the true age groups of participants is coded as an ordinal variable in the following fashion –1: "0-11", 2: "12--17", 3: "18-24", 4: "25-34", 5: "35-54", 6: "55-74", 7: "75+".

As an additional robustness test, we examine whether the relationship (slope) between the true age of the target face and subjects' perceptions of this face's age varies by the true gender of the target face. This relationship is examined through an OLS regression that predicts the perceived age of the target face as a function of an interaction between the true gender and age of the face (i.e., true age x true gender). This model is presented in table S5, while also including fixed effects to control for

participant, the race of the target face, and whether a participant is familiar with the face. We do not find a significant effect of an image's true gender on the relationship between an image's true age and a subject's perception of this image's age ($\beta[\text{True Age} \times \text{Male}] = 0.00$, $\text{CI} = [-0.01, 0.01]$, $p = 0.97$). In fact, the effect of this interaction is essentially zero, though expectedly, there remains a very strong positive association between the true age of a face and a subjects' perception of this faces age ($\beta[\text{True Age}] = 0.91$, $\text{CI} = [0.90, 0.92]$, $p = 2.2 \times 10^{-16}$). Overall, this model captures an impressive amount of variance in the predicted outcomes ($R^2 = 0.91$). This model lends further evidence to the claim that participants' perceptions of people's age did not vary as a function of their gender.

The performance of human coders in our sample is higher than the accuracy of human age judgments identified by prior work, which found that humans could only accurately identify the correct age group of faces 60% of the time³; however, this prior work did not include any faces below 20 years old and above 70 years old, thereby excluding age groups that are essential to the evaluation of the age-related gender bias explored in our study. Moreover, this prior work only examined the rate at which human coders identified the precise age group of particular faces. However, since our work is also interested in differences in the distribution of age associated with men and women in online images, we also evaluate the extent to which coders' judgments of age fall close together along the age continuum. Adopting this approach reveals significantly higher levels of accuracy and coherence in coders' age judgments across genders and age groups. As described above, the coders exhibit substantial coherence in their classifications – with nearly all judgments falling within one age group of each other – lending further support to the quality of our sample.

B. 1. 5 Evaluating Intercode Agreement in Gender and Age Classification

Another important indicator of data quality in crowdsourced human judgments is intercode agreement. We follow prior work on stereotypes in Google Images by hiring three unique human coders to classify each face^{4,5}. Every face from every search query was uniquely uploaded for classification. As a consequence, images that repeated across questions were uploaded and classified by more than 3 coders. 11% of images in the Google data repeated across searches, and 4% of the images in the Wikipedia data repeated (both across categories, and across the explicitly gendered and non-explicitly gendered searches of the same category in the case of Google). As a result, these repeated images were associated with more than 3 coders; on average, they were classified by 5 unique coders ($\text{SD} = 1.8$), with a maximum of 36 coders. All results equally hold when controlling for repeat images.

Across all images, coders reached unanimous agreement in their gender and classifications for 64% of images. Related supplementary analyses show that all of our results equally hold if we only examine images associated with consensus gender and age classifications (Fig. S3). The rate of intercode agreement is expectedly lower for age classifications since age was indicated using a greater variety of options. Importantly, since age classifications can be ranked along an ordinal dimension from youngest to oldest, we are able to determine whether subjects generally agreed on where to classify the age of a face along the age continuum. Indeed, over 90% of all age classifications for the same face belonged to an adjacent age group (i.e., within one age group of each other). This measure indicates that our coders agreed qualitatively almost all of the time in terms of whether a face belonged on the lower or upper part of the age continuum.

We also show that intercode agreement similarly holds when controlling for the gender of the coders in our sample. In general, female coders classified 52.2% of images as female, and male coders classified 48.1% of images as female. While this distribution suggests that coders are slightly biased toward classifying faces in accord with their own gender, this effect is strikingly weak; both male and female coders classified faces as male or female at a roughly 50/50 rate. Most importantly, section A. 1. 3 of this appendix shows that all of our main results are highly robust

to controlling for the gender and age of the coder, so these patterns do not provide a plausible confounding variable when accounting for our core findings.

For additional robustness, we also calculate the chance-corrected intercoder reliability of raters in our sample using GWET's Agreement Coefficient (AC)⁶. We calculated GWET's AC using the irrCAC package in R. As reported in prior work, the coders in our sample reached a GWET's coefficient of 0.48 for their gender classification and 0.41 for their age classification. The GWET coefficient for both measures falls within the 'good' range of reliability according to standard interpretations of this measure, especially considering that our sample was limited to only three coders per image. Combined with our percent-agreement results, these analyses suggest that our coders provided reliable statistical judgments of the gender of faces in our sample.

B. 1. 6 Sensitivity Analyses Demonstrating Robustness to Potential Coder Biases

A potential concern is annotator bias – in particular, we seek to rule out the concern that annotators are biased toward underpredicting the age of younger women and overpredicting the age of older women, since a bias of this kind could give rise to our results despite no underlying difference in the true age of faces. While our validation experiment does not provide an indication of this bias in our coder sample, as an additional robustness test, we use sensitivity analyses to show that our main results continue to hold even if we assume and control for a bias of this kind among our coders. For these sensitivity analyses, we assume that annotators underpredicted the age of all images of young females in our entire Google dataset; we then ask, for how many of these images would we have to 'correct' this bias to eliminate the difference between the age distribution of female and male faces in our dataset. In this simulation, a fixed fraction of young female images from age groups 1, 2, and 3 is selected, and then the age group assigned to these selected images is increased by 1 (to adjust for underprediction, e.g., going from 12-17 to 18-24).

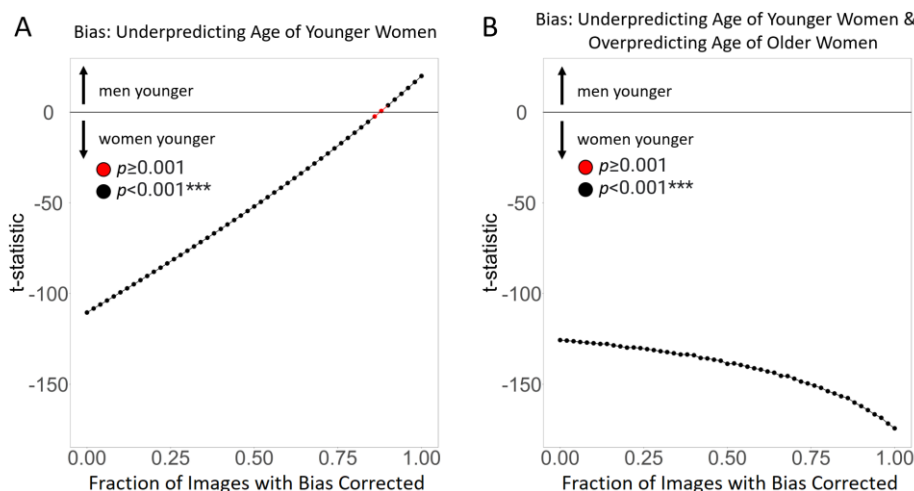


Figure S6. Evaluating the statistical significance of the difference in age distribution between the female and male faces in our entire Google Image dataset, while altering the fraction of images for which a hypothetical annotator bias is corrected. Panel A shows the results of this simulation when correcting the hypothesis bias of annotators to underpredict the age of younger women; this bias is corrected by randomly selecting a fixed fraction of images of younger women (age groups 1, 2, and 3) and increasing their assigned age by 1. Panel B shows the same results as panel A, while also including the effects of correcting the hypothetical bias of annotators to overpredict the age of older women; in this simulation, a fixed fraction of

images is selected, and if they are younger (belonging to age groups 1, 2, and 3), then their age group is increased by 1, and if they are older (belonging to age groups 5, 6, and 7), then their age group is decreased by 1. Black dots indicate statistically significant differences at the $p < .001$ level; red dots indicate a statistical difference above the $p > .001$ level.

Panel A of figure S6 shows how the fraction of ‘corrected’ images changes the statistical difference in the age distribution of male and female faces. We find that over 75% of young female images would need their age group to be increased by 1 for there to no longer be a significant age difference between male and female faces in our Google dataset. This means that the subjective bias to underpredict the age of younger women would have to be remarkably strong for it to pose an issue for our results. Overall, this seems implausible, given that we find no evidence of such bias in our validation task (see section A. 1. 4).

Panel B of figure S6 shows the same simulation as panel A, while also including the effects of correcting the additional bias, whereby annotators also overpredict the age of older women. In this simulation, a fixed fraction of female images is selected, and if they are younger (belonging to age groups 1, 2, and 3), then their age group is increased by 1 (as in Panel A), and if they are older (belonging to age groups 5, 6, and 7), then their age group is decreased by 1 (adjusting the overprediction). We see that, even when correcting for the hypothetical bias of annotators to underpredict the age of young women and overpredict the age of older women, our main results still equally hold, suggesting that even if these hypothetical biases were to play, it is highly unlikely that they would undermine our results (note these outcomes are even stronger if we examine only the bias to overpredict the age of older women, which would simply make the faces of women in our sample even younger). These results equally hold in our Wikipedia dataset.

Variables	Beta	95% CI ¹	p-value
Gender of Image			
Female	—	—	—
Male	0.37	0.37, 0.38	0.000000e+00
Data Source			
Google	—	—	—
Wikipedia	-0.12	-0.13, -0.11	1.166783e-85
Gendered Category	0.01	0.00, 0.01	1.373656e-01
Age Category	-0.25	-0.26, -0.25	0.000000e+00
Google Image U.S. Search Freq	-0.56	-0.58, -0.54	0.000000e+00
Polysemy	-0.01	-0.01, 0.00	4.005226e-24
Word Frequency Scaled	0.02	0.02, 0.02	1.205347e-114
<i>Statistic</i>	9,509		
R^2	0.030		
<i>Adjusted R²</i>	0.030		
<i>No. Obs.</i>	2,121,664		
<i>df</i>	7		
<i>AIC</i>	6,802,828		

Variables	Beta	95% CI ¹	p-value
<i>Residual df</i>	2,121,656		
<i>Sigma</i>	1.20		

¹CI = Confidence Interval

Table S6. An OLS regression predicting coders' age classification of faces, which are treated as an ordinal numeric variable coded in the following way: (1) 0-11, (2) 12-17, (3) 18-24, (4) 25-34, (5) 35-54, (6) 55-74, (7) 75+. This model examines the correlation between this outcome variable and coders' gender classification, while controlling for a range of linguistic features for each social category, including their gender connotation, age connotation, their usage frequency (both in general and in Google Image search specifically), and their polysemy.

B. 1. 7 Robustness to Controlling for Linguistic Features of Social Categories

Here, we demonstrate the robustness of our main results to controlling for a range of linguistic features at the level of social categories. Table S6 presents an OLS model predicting coders' age classifications of faces, while controlling for a range of linguistic features for each social category, including their gender and age connotation, their usage frequency (both in general language use and in Google Image search specifically), and their polysemy. The model shows that, subject to all of these controls, images classified as male are strongly associated with significantly higher age classifications ($\beta[\text{Male}] = 0.37$, $\text{CI} = [0.37, 0.38]$, $p = 2.2 \times 10^{-16}$). While each of these linguistic features is correlated significantly with the outcome, the combined, they account for strikingly little variation in age predictions; the same model without any of these linguistic features captures an R^2 of 0.024, and adding these linguistic features only increases R^2 to 0.03, marking only a minor gain to the model's predictive power.

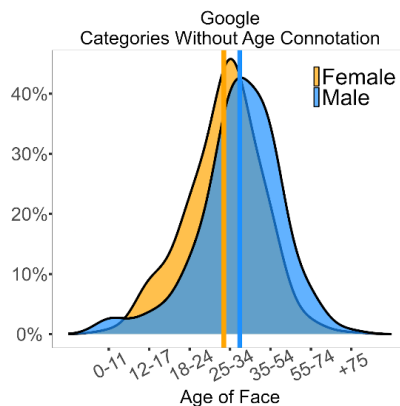


Figure S7. Women are represented as significantly younger than men in a large sample of images ($n = 159,668$) from Google associated with all social categories in Wordnet, while excluding categories that connote age, yielding 3,034 categories). Gold (blue) lines indicate the average age for female (male) faces.

We conclude by showing that our main results are robust to only examining Google Images (searched without specifying gender) associated with categories that lack explicit age connotations. Examples of categories with explicit age connotations include 'child' and 'adult.' We would like to note that we took a conservative approach by excluding all categories that moderately implied age

(e.g., ‘beginner’ and ‘student’), even though such associations do not always indicate age (older people can still be beginners as a function of context). Even while conservatively excluding age-related categories, women continue to be presented as significantly younger than men in Google images ($t = -75.03$, $p = 2.2 \times 10^{-16}$, $n = 3,034$ categories, 136,023 images; Fig. S7).

B. 1. 8 Robustness to Image Features

Here, we show the robustness of our main results to controlling for a range of image features. Table S4 presents an OLS model predicting coders’ age classifications of faces, while controlling for a range of image features, including the number of faces in each image, the number of images per category, the number of faces per category, the number of social categories that each image is associated with (e.g., the number of different google search terms that returned a given image; in other words, repeats), and whether the face shows evidence of being a digital avatar rather than a photograph (18% of images were flagged as avatars by our coders). The results are shown in Table S7.

Variables	Beta	95% CI ¹	p-value
Gender of Image (Mode)			
Female	—	—	—
Male	0.25	0.25, 0.26	0.000000e+00
Data Source			
Google	—	—	—
Wikipedia	-0.24	-0.25, -0.22	5.278324e-156
Avatar			
No	—	—	—
Yes	-0.64	-0.65, -0.64	0.000000e+00
Num. Faces per Image	0.00	0.00, 0.00	6.144069e-01
Num. Faces per Categ.	0.00	0.00, 0.00	2.239100e-33
Num. Images per Categ.	0.00	0.00, 0.00	0.000000e+00
Num. Categ. per Image	0.00	0.00, 0.00	9.401224e-01
<i>Statistic</i>	<i>9,002</i>		
<i>R²</i>	<i>0.069</i>		
<i>Adjusted R²</i>	<i>0.069</i>		
<i>No. Obs.</i>	<i>848,101</i>		
<i>df</i>	<i>7</i>		
<i>AIC</i>	<i>2,487,802</i>		
<i>Residual df</i>	<i>848,093</i>		
<i>Sigma</i>	<i>1.05</i>		

¹CI = Confidence Interval

Table S7. An OLS regression predicting coders’ age classification of faces, which are treated as an ordinal numeric variable coded in the following way: (1) 0-11, (2) 12-17, (3) 18-24, (4) 25-34, (5) 35-54, (6) 55-74, (7) 75+. This model examines the correlation between this

outcome variable and coders' gender classification, while controlling for a range of image features.

Table S7 shows that, controlling for image features, images classified as male are still strongly associated with significantly higher age classifications ($\beta = 0.25$, $CI = [0.25, 0.26]$, $p = 2.2 \times 10^{-16}$). There was no significant relationship between the number of search terms associated with an image and coders' age judgments ($p = 0.94$), nor between the number of faces in an image and coders' age judgments ($p = 0.61$). The number of faces per category and the number of images per category are both very weakly but yet significantly positively associated with higher age estimates (at the $p < 0.00001$ level). Avatars are associated with significantly lower age judgments ($\beta = -0.64$, $CI = [-0.65, -0.64]$, $p = 2.2 \times 10^{-16}$).

B. 1. 9 Robustness to Non-cropped Image Classification

Here, we show that our results are robust to whether or not the faces in each image are automatically cropped prior to their classification by human coders. For this robustness test, we asked a new sample of coders from Mechanical Turk to classify the age and gender of the focal faces in the uncropped Google images that we originally collected as part of our main study.

Specifically, coders in this task were given two criteria when classifying the uncropped images: they were asked to focus on (i) the "focal face" of the person who (ii) belongs to the search category used to retrieve the image (e.g., "doctor"), following the standard methodology employed by recent crowdsourcing studies of stereotypes in online images^{4,5}. For this task, we recruited a separate sample of 1,004 coders to classify the focal faces in all of the original, uncropped Google images associated with 300 categories randomly sampled from the broader set of 3,434 categories presented in the main text. Our preprocessing and data preparation procedures were otherwise identical to that of the main study; that is, we removed the judgments of all human coders who failed to complete a simple attention check, and we excluded all images that were identified as failing to display a human face. This approach resulted in 26,529 images.

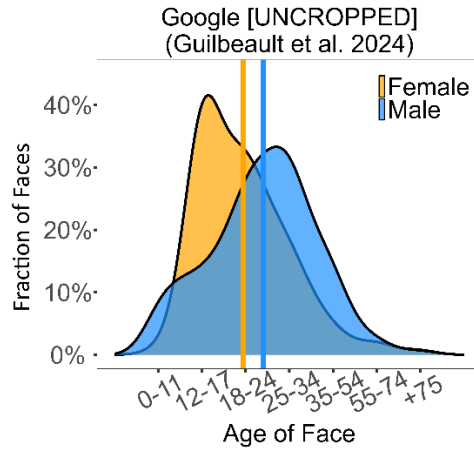


Figure S8. Human judgments of age for uncropped images of men and women across (A) 300 social categories in Google Images (randomly sampled from the broader set of 3,434 categories presented in the main manuscript). This sample includes $n = 26,529$ images in total. Solid gold (blue) lines indicate the modal age for female (male) faces.

Within this uncropped sample, we find that the same patterns of gendered ageism hold with equal levels of statistical significance (Fig. S8). Specifically, the modal age for women is 12-17 (adolescence), whereas the modal age for men in this sample is 25-34 (young adulthood). The average age of women is significantly lower than the average age of men in this sample at high statistical significance ($t = -29.21$, $p = 2.2 \times 10^{-16}$, Student t-test, two-sided). These results further support our argument that the correlation between age and gender in our main results is not driven by idiosyncrasies or biases in the facial cropping algorithm we used.

Despite the robustness of our results to the additional data and supplementary analyses described above, survey-based crowdsourcing methods are limited in terms of their ability to once-and-for-all rule out the influence of annotators' subjective biases on the classifications they produce. For example, in the context of analyzing uncropped images, we adopt the methodology employed by prior crowdsourcing studies of stereotypes in online images, which relies on annotators to decide which face is focal for them in a given image, based on their subjective judgment^{4,5}. However, while this introduces subjective bias into our classification data for this analysis, our primary interest is in studying the faces that people encounter online, and therefore, we are primarily concerned with whichever faces coders deemed to be focal according to them, since this indicates which faces they were most likely to observe in each image. It also helps that participants in this task were asked to focus on the focal face corresponding to the relevant search category used to retrieve the image, which further constrains their selection process and promotes convergent judgments across coders; for example, by steering coders toward the identification of the "doctor" in an image of a doctor and a patient, where both are equally focal. The topic of which faces appear to be salient in images – both to humans and computer vision algorithms⁷ – remains a growing area of research, and advancements in this direction stand to further enrich the crowdsourcing techniques leveraged in this robustness analysis.

B. 1. 10 Validation of the OpenCV Image Cropping Algorithm

A potential concern is the extent to which the OpenCV face-cropping algorithm used in this study is accurate and reliable, especially in light of prior work demonstrating demographic-related biases in popular face detection algorithms⁸⁻¹⁰. To ensure that our use of the OpenCV cropping algorithm did not introduce confounding variables due to gender-related biases in its cropping functionality, we examined the rate of false negatives (i.e., the rate at which the OpenCV algorithm missed faces in images) in a random sample of 1,000 randomly selected images from our dataset. We confirm that the rate at which the OpenCV algorithm missed faces in images (i.e., the false negative rate) was fairly low (<8%), and manual inspection reveals that the majority of false negatives concern images where the faces are either too blurry or small to observe, or where the faces are occluded by headwear (e.g., helmets) or other objects, all of which are conditions where it is challenging for humans and algorithms to identify faces, let alone their age and gender. For this same reason, these are the faces our hypotheses are least concerned with, since our goal is to characterize patterns of gendered ageism in those faces that are salient, interpretable, and most likely to be encountered by internet users. Most importantly, we observe that the false negative rate was equally conserved across both male and female faces. Among the faces that were missed and could be clearly identified, 48% were male and 52% were female, exhibiting no significant difference ($p = 0.27$, Proportion Test). Moreover, we do not observe any significant difference in the age of the female faces missed by our algorithm and the age of the female faces that were detected by our algorithm ($p = 0.45$, Student t-test, two-tailed); nor do we observe any significant difference in the age of the male faces missed by our algorithm and the age of the male faces that were detected by our algorithm ($p = 0.37$, Student t-test, two-tailed). This false negative rate is comparable to the performance of the OpenCV algorithm as measured in prior studies¹¹⁻¹³.

We go further by using sensitivity simulations to show that even if the false negative faces completely reversed our predictions, the robustness of our main effects continues to hold. For this sensitivity analysis, we simulate the effects of adding the false negative faces to our sample, while assuming that the missed male faces belong to the younger age groups (randomly sampled from groups 1, 2, or 3) and that the missed female faces belong to the older age groups (randomly sampled from groups 5, 6, or 7). Consistent with the above validation test, this simulation assumes an 8% false negative rate, including 48% men and 52% women. To correct for this false negative rate in our main Google dataset of 159,668 images (i.e., those collected without specifying gender, as in “male doctor”), we added 6,131 young male faces with randomly assigned categories and 6,642 female faces with randomly assigned categories. To correct for this false negative rate in our Wikipedia dataset of 14,312 images, we added 550 young male faces with randomly assigned categories and 595 female faces with randomly assigned categories.



Figure S9. Results of a sensitivity test that adds 8% of faces into the sample (to compensate for the false negative face cropping rate), while assuming that all male faces added are younger and all female faces added are older (contrary to our hypothesis). (A) Distribution of ages by gender in the main Google dataset ($n = 159,668$ images) with 6,131 young male faces and 6,642 older female faces added; (B) distribution of ages by gender in the Wikipedia dataset ($n = 14,312$ images) with 550 young male faces and 595 older female faces added.

Even when including the faces missed by our algorithm assuming the opposite pattern to our main results, we continue to find that women in our Google dataset are represented as significantly younger than men ($t = -8.28$, $CI = [-0.05, -0.03]$, $p = 2.2 \times 10^{-16}$, Student t-test, two-tailed), as shown in Figure S9A. Similar patterns of robustness replicate in our Wikipedia data. Women continue to appear significantly younger than men in our Wikipedia dataset, even when including the biased false negative faces ($t = -14.93$, $CI = [-0.36, -0.28]$, $p = 2.2 \times 10^{-16}$, Student t-test, two-tailed), as shown in figure S9B.

This sensitivity analysis is an especially conservative test of the robustness of our results, since it assumes that these false negative faces are all clearly recognizable and non-occluded (which is often not the case), *and* that their age distribution is in the opposite direction of our theory, and yet still we show that it does not change our main results, given the robustness of the large-scale trends we identify. Note that the dynamics of false negatives in our face cropping algorithm are not relevant to our findings using the ground-truth IMDB and Wikipedia datasets, since the image classifications in these datasets do not depend on the use of our face cropping algorithm.

We also control for OpenCV’s false positive rate, i.e., the rate at which it misidentified a non-face object as a human face. We asked each annotator to identify whether each image they classified displayed a human face. The annotators judgments indicate that 18% of the cropped images were false positives and could not be reliably identified as a human face. As described in the main text, we

control for false positives by removing all images for which at least one human coder identified the image as a false positive. Note, all of our main results equally hold if no images and no classification judgments are removed from our analyses.

B. 1. 11 Sources of Images in Google Image Dataset

Here, we analyze the kinds of websites from which the images in our Google dataset derive. Note, this analysis and figure is reproduced from Guilbeault et al. (2024) with the authors' permission¹⁴. Google largely serves as an expeditor for routing users to content that is hosted on other websites. This raises the question of which kinds of websites are providing the images that appear in our Google image dataset. If most of the images in our dataset derived from stock photography websites, for instance, this would significantly inform the interpretation of our results, since the age-related gender biases we observe would then be more appropriately described as a result of marketplace dynamics over stock photography websites, rather than as a more general bias that likely appears in images across a variety of sources. For this reason, we classify the types of websites from which our Google image dataset derives. To date, there is no publicly available and standardized method for automatically categorizing websites. It remains a highly manual task, since it involves visiting websites and observing various qualitative features. As a result, we were unable to classify all of the websites linked in our dataset, which approach half a million; but for supplementary purposes, we hired a team of three undergraduate assistants to manually categorize the websites associated with a random subset of 2,000 images from our dataset (randomly sampled, without replacement, from any category). We adopted a labeling scheme that poses *prima facie* validity. In a number of cases, websites could be assigned to multiple categories; in such cases, we asked our undergraduate coders to highlight the most fitting category, which is what we used. The coders were asked to review each other's labels and come to a consensus on the categorization of each website. The distribution of images by website type is provided below.

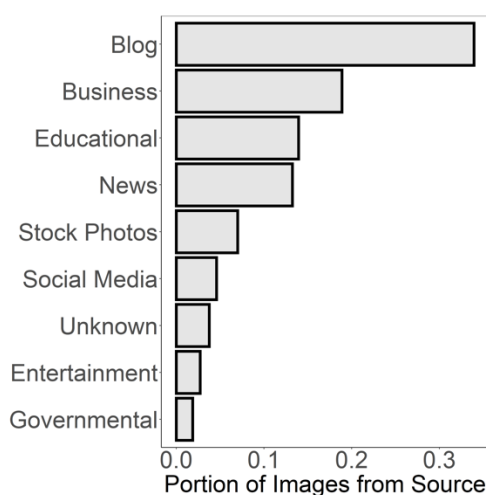


Figure S10. The proportion of images in our Google Image dataset deriving from each type of website. The total portion of images sums to 1. These results are based on a random sample of $n = 2000$ images from our dataset.

Figure S10 presents the proportion of images in our Google Image dataset deriving from each type of website, according to the manual classifications of a team of undergraduate assistants

applied to a random subset of 2,000 images. The results show that the images in our Google Image dataset derived from a variety of sources, such that no particular kind of website sourced the majority of images. Images most commonly derived from personal blogs or business websites, followed closely by educational, news, and stock photography websites. This suggests that the age-related gender biases we observe are driven in part by content that internet users have elected to display on their personal websites, but also by market dynamics relating to audiences' preferences when choosing which news to consume or which images to purchase. More generally, this suggests that our findings are likely to characterize online images from a diversity of sources, where these images are produced and circulated via a range of mechanisms and for a range of purposes. We discuss the opportunities for future research that these analyses reveal in the discussion section of our main study.

B. 1. 12 Benchmarking Image Data Against the U.S. Census

Here, we test whether the aggregated age judgments of faces provided by human coders in our main Google and Wikipedia image datasets are consistent with the ground truth sociodemographic distributions of age across occupations according to the U.S. Census. We were able to match 867 categories from these image datasets to occupational categories in the U.S. Census. We focus on comparing against Census data from a relevant time frame, specifically from 2020 to 2023 (where the 2023 Census record is the most recent to date). We then compare whether the average age rating (across the seven ordinal age categories; see “Materials and Methods”) across all faces for each occupation (i.e., those judgments provided by coders from Mturk) correlate with the median age of people in the same occupations according to the U.S. Census. Note, the correlation between (the average of) participants' ordinal age judgments of online faces and the median age of people in organizations from a nationally representative sample is *not* expected to be particularly high, given the many foundational differences in how age is measured and sampled across these different data structures. In this case, we are interested in observing whether there is any reasonable and significant correlation in the expected direction, where even a moderately strong correlation is considered impressive under this conservative analysis.

Despite these constraints due to the different formatting of each respective data structure, we nevertheless uncover broad stroke correlations which show that the Mturkers' judgments of which faces are younger vs. older are indeed predictive of differences in the median age of people across occupations in the Census.

Variables	Beta	95% CI ¹	p-value
Avg. Age Rating of Face in Image	1.03	0.31, 1.77	0.0053
Census year			—
2020	—	—	—
2021	-0.19	-0.69, 0.31	0.015812
2022	0.10	-0.39, 0.60	0.216884
2023	-0.27	-0.77, 0.22	0.011701
Statistic	7.10		—
R ²	0.010		—

Variables	Beta	95% CI ¹	p-value
Adjusted R ²	0.009		—
No. Obs.	2,822		—
df	4		—
AIC	16,816		—
Residual df	2,817		—
Sigma	4.76		—

¹CI = Confidence Interval

Table S8. An OLS regression predicting the median age of people in a given occupation according to the U.S. Census (2020-2023) as a function of the average age estimate of the faces associated with each occupation in the observational sample of Google images, while controlling for fixed effects by Census year, and while clustering standard errors at the level of occupational category.

Table S8 presents the results of an OLS regression predicting the median age of people in a given occupation as a function of the average age estimate of the faces associated with each occupation in the observational sample of Google images, while controlling for fixed effects by Census year, and while clustering standard errors at the level of occupational category. The coders' age judgments for each occupation are averaged across genders. The results show that the average age associated with an occupation in Google images is positively and significantly correlated with the median age of people in this occupation according to the U.S. Census ($\beta[\text{Avg. Age Rating of Face in Image}] = 1.03$, $\text{CI} = [0.31, 1.77]$, $p = 0.005$), holding the Census year constant. For reference, the raw Pearson correlation between average age coded in Google images and the median Census age of an occupation is $r = 0.10$ ($p = 4.515 \times 10^{-7}$, two-tailed). Replicating Table S8 while including the image data from Wikipedia – and while controlling for whether the images derive from Google or Wikipedia – retains the significant effect ($\beta[\text{Avg. Age Rating of Face in Image}] = 0.47$, $\text{CI} = [0.21, 0.74]$, $p = 0.0004$), with no significant effect of data source ($\beta[\text{Google} | \text{Wikipedia}] = -0.15$, $\text{CI} = [-0.50, 0.19]$, $p = 0.39$), but clustering standard errors by occupation in this model causes the main correlation to fall just beyond traditional bounds of significance ($\beta[\text{Avg. Age Rating of Face in Image}] = 0.47$, $\text{CI} = [-0.05, 1.03]$, $p = 0.07$), suggesting that the correlation in the Wikipedia is substantially less salient, which is not unsurprising, given that the number of images available per occupation in the Wikipedia dataset is order of magnitudes less than the number available in the Google dataset.

Next, we examine whether the representation of gender across occupations in each of these datasets is significantly associated with the median age of people belonging to these occupations according to the Census, as a potential indicator of latent age-based gender bias. For each occupation in both the Google and Wikipedia dataset, we identify the fraction of faces identified as Male. This is consistent with prior studies of this data, which used a similar analytic procedure to identify the statistical gender association of social categories in online images¹⁴. We then determine whether the extent of male bias in the images associated with each occupation is correlated with its median age in the U.S. Census from 2020 to 2023. The results of this model are presented in Table S9, below.

Variables	Beta	95% CI ¹	p-value
P(Male Face)	2.8	0.39, 5.21	0.02241
Census year			—
2020	—	—	—
2021	-0.19	-0.35, -0.02	0.03087
2022	0.13	-0.04, 0.29	0.12926
2023	-0.27	-0.48, -0.06	0.01103
Data Source			—
Google	—	—	—
Wikipedia	-0.65	-1.38, 0.08	0.08110
Statistic	6.56		—
R ²	0.009		—
Adjusted R ²	0.007		—
No. Obs.	3,808		—
df	5		—
AIC	22,733		—
Residual df	3,802		—
Sigma	4.78		—

¹CI = Confidence Interval

Table S9. An OLS regression predicting the median age of people in a given occupation according to the U.S. Census (2020-2023) as a function of the fraction of male faces associated with this occupation in online images from Google and Wikipedia, while controlling for fixed effects by data source (Google or Wikipedia), Census year, and while clustering standard errors at the level of occupational category.

Table S9 presents the results of an OLS regression predicting the median age of people in a given occupation according to the U.S. Census (2020-2023) as a function of the fraction of male faces associated with this occupation in online images from Google and Wikipedia, while controlling for fixed effects by Data Source (Google or Wikipedia), Census year, and while clustering standard errors at the level of an occupational category. The model shows that the probability of observing a male face in the online images associated with an occupation is positively and significantly correlated with the median age of people in the same occupation according to the U.S. Census ($\beta[P(\text{Male Face})] = 2.8$, CI = [0.39, 5.21], $p = 0.02$), holding data source (Google or Wikipedia) and Census year constant. For reference, the raw correlation between the probability of a male face in the Google Images for a given occupation and this occupation's Census age is $r = 0.07$ ($p = 0.0004$, $t = 3.54$, Pearson correlation, two-tailed); and the raw correlation between the probability of a male face in the Wikipedia Images for a given occupation and this occupation's Census age is $r = 0.13$ ($p = 6.718 \times 10^{-5}$, $t = 4.00$, Pearson correlation, two-tailed). Together, these findings indicate that the extent of male bias in the Google and Wikipedia images associated with a given occupation is predictive of the median age estimate of


people in this occupation in the U.S. Census, providing yet another indicator of latent age-based gender bias anchored in ground truth sociodemographic data of real-world age distributions across occupations.

Think of this ladder as where people in society stand.

At the top of the ladder are people who are the best off - those who have the most money, best education, and most respected jobs.

At the bottom of the ladder are the people who are the worst off - those who have the least money, least education, and the least respected jobs.

The higher up you are on this ladder, the closer you are to the people at the top; the lower you are, the closer you are to the people at the bottom.



Imagine someone who belongs to the following occupation: `{lm://Field/1}`. **With only this information, where do you think this person is placed on the ladder above?**

1 2 3 4 5 6 7 8 9 10

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

Figure S11. A screenshot of the language used for the General Social Survey (GSS) ladder question designed to elicit participants’ perceptions of the status, prestige, and income associated with occupations. The bold “`{lm://Field/1}`” text indicates where the name of the randomly sampled occupation is piped in from the full list of occupations examined ($n = 867$ occupations).

B. 1. 13 Robustness of Status Associations in Google Images

Here, we demonstrate the robustness of our results linking age-gender associations in Google images with human perceptions of occupational status and the yearly earnings of occupations according to the U.S. We begin by discussing how our main results involving human perceptions of status (Panel A of Extended Data Figure 4) replicate for both the status and prestige questions separately (our main results aggregated across these questions), as well as when using the standard GSS social status ladder question. While the GSS social status ladder question has been widely used in the past, we do not use it in our main results because recent work shows that people’s responses to this question are highly variable and heterogeneous. In addition, this question asks participants to consider many aspects of occupations simultaneously, including status, prestige, respectability, education and income, so it is less clear which features are driving participants’ associations (see Fig. S11). Nevertheless, as we report below, the same patterns replicate when using the GSS ladder question as well.

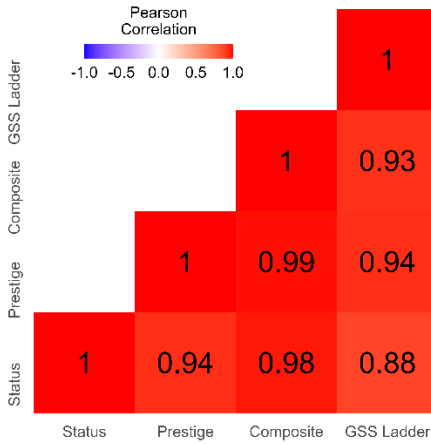


Figure S12. Heatmap displaying the pairwise correlations between the different measures of perceived social status of occupations ($n = 867$ occupations): (i) the status indicator (How would you rate the social status of someone belonging to this occupation? -2: very negative, -1: negative, 0: neutral, 1: positive, 2: very positive); (ii) the prestige indicator (To what extent do you agree that it is prestigious to belong to this occupation? -2: strongly disagree, -1: disagree, 0: neutral, 1: agree, 2:strongly agree); (iii) the composite measuring based on averaging the status and prestige indicator (shown in panel A of Extended Data Figure 3); and (iv) participants' responses to the GSS social status question, which asks participants to position occupations on a ladder of social status consisting of 10 rungs, where the bottom rung refers to occupations with the lowest status, prestige, and earnings, and the top rung refers to occupations with the highest status, prestige, and earnings. All correlations indicate pairwise Pearson correlations, two-tailed.

First, we emphasize that participants' responses to all of these questions are highly correlated with each other at the occupational level (Fig. S12). In addition, we confirm that the probability of men being older than women in Google images (represented as a binary variable, 1 = men as older) is independently and separately predicted by the status indicator ($r = 0.06$, $t = 9.09$, $CI = [0.05, 0.07]$, $p = 2.2 \times 10^{-16}$) and the prestige indicator ($r = 0.09$, $t = 13.33$, $CI = [0.07, 0.10]$, $p = 2.2 \times 10^{-16}$), (Pearson correlation, two-tailed, $N = 22,922$ responses across 866 occupations). We proceed to demonstrate the robustness of our results by examining the composite measure and the GSS ladder question.

Variables	Beta	95% CI ¹	p-value
Status + Prestige (Composite Measure)	0.04	0.03, 0.04	0.00067
Participant Fixed Effects Included	—	—	—
<i>Statistic</i>	1.12		—
R^2	0.050		—
<i>Adjusted R²</i>	0.005		—
<i>No. Obs.</i>	22,399		—

Variables	Beta	95% CI ¹	p-value
<i>df</i>	1,002		—
<i>AIC</i>	22,866		—
<i>Residual df</i>	21,396		—
<i>Sigma</i>	0.394		—

¹CI = Confidence Interval

Table S10. An OLS regression predicting the probability of men being older than women in Google images of an occupation (binary outcome, 1 = men as older) as a function of participants' ratings of each occupation's status and prestige (averaged into a composite measure). This model controls for participant fixed effects ($N = 1,002$ participants), while also clustering standard errors at the occupation level. Participants were recruited using Prolific's nationally representative U.S. sample.

First, we demonstrate the statistical robustness of the positive correlation between participants' ratings of each occupation's status and prestige (averaged into a composite measure) and the probability of men appearing older than women in Google Images of each occupation (binary outcome, 1 = men as older). This model controls for participant fixed effects ($N = 1,002$ participants), while also clustering standard errors at the occupation level. Table S10 shows that, subject to these controls, participants' status and prestige judgments continue to be significantly and positively correlated with the probability of men being older than women in corresponding Google Images of occupations ($\beta[\text{Composite Measure}] = 0.04$, $\text{CI} = [0.03, 0.04]$, $p = 0.0006$).

Variables	Beta	95% CI ¹	p-value
GSS Ladder Social Status Indicator	0.02	0.02, 0.03	2.304e-05
Participant Fixed Effects Included	—	—	—
<i>Statistic</i>	1.32		—
<i>R²</i>	0.060		—
<i>Adjusted R²</i>	0.015		—
<i>No. Obs.</i>	22,922		—
<i>df</i>	1,002		—
<i>AIC</i>	23,227		—
<i>Residual df</i>	21,919		—
<i>Sigma</i>	0.393		—

¹CI = Confidence Interval

Table S11. An OLS regression predicting the probability of men being older than women in Google images of an occupation (binary outcome, 1 = men as older) as a function of participants' ratings of each occupation's status using the GSS ladder social status indicator (see Fig. S8). This model controls for participant fixed effects ($N = 1,002$ participants), while

also clustering standard errors at the occupation level. Participants were recruited using Prolific's nationally representative U.S. sample.

Next, we demonstrate the statistical robustness of the positive correlation between participants' ratings of each occupation's status and prestige using the GSS ladder question and the probability of men appearing older than women in Google Images of each occupation (binary outcome, 1 = men as older). This model controls for participant fixed effects ($N = 1,002$ participants), while also clustering standard errors by occupation. Table S11 shows that, subject to these controls, participants' status associations based on the GSS ladder indicator continue to be significantly and positively correlated with the probability of men being older than women in corresponding Google Images of occupations ($\beta[\text{Composite Measure}] = 0.02$, $CI = [0.02, 0.03]$, $p = 2.2 \times 10^{-5}$).

Variables	Beta	95% CI ¹	p-value
Earnings (Logged)	0.08	0.06, 0.11	0.0120
Error of Earnings (Logged)	0.00	-0.02, 0.01	0.8569
Census Year			—
2015	—	—	—
2016	0.00	-0.05, 0.05	0.2710
2017	0.00	-0.05, 0.05	0.3055
2018	0.02	-0.03, 0.06	0.0954
2019	0.02	-0.03, 0.06	0.1375
2021	0.01	-0.04, 0.06	0.4286
2022	0.01	-0.04, 0.05	0.6640
<i>Statistic</i>	5.27		—
<i>R²</i>	0.009		—
<i>Adjusted R²</i>	0.008		—
<i>No. Obs.</i>	4,523		—
<i>df</i>	8		—
<i>AIC</i>	4,633		—
<i>Residual df</i>	4,514		—
<i>Sigma</i>	0.403		—

¹CI = Confidence Interval

Table S12. An OLS regression predicting the probability of men being older than women in Google images of an occupation (binary outcome, 1 = men as older) as a function of the average yearly earnings (logged) associated with each occupation in the U.S. Census. This model controls for (i) the logged error of the median yearly earnings (provided by the Census) and (ii) the year of the Census data examined (from 2015 to 2022; note: the 2020

COVID year is missing in the Census data provided). Standard errors are clustered at the occupation level.

Next, we show that the correlation between the yearly earnings¹ of an occupation and the probability of men appearing older than women in Google Images of occupations is robust to statistical controls. Table S12 presents an OLS regression predicting the probability of men being older than women in Google images of an occupation (binary outcome, 1 = men as older) as a function of the average yearly earnings (logged) associated with each occupation in the Census, while controlling for (i) the logged error of the median yearly earnings (provided by the census) and (ii) the year of the Census data examined, and while also clustering standard errors at the occupation level. Table S9 shows that, subject to all of these controls, the logged yearly earnings of an occupation continue to be significantly predictive of the probability of men being older than women in corresponding Google Images of occupations ($\beta[\text{Earnings Logged}] = 0.08$, $CI = [0.06, 0.11]$, $p = 0.01$).

Variables	Beta	95% CI ¹	p-value
Gender Pay Gap (Men - Women)	0.00	0.00, 0.00	0.0179807
Error of Earnings (Women)	0.00	0.00, 0.00	0.0003435
Error of Earnings (Men)	0.00	0.00, 0.00	0.0088706
Census Year			—
2015	—	—	—
2016	0.00	-0.07, 0.06	0.1145227
2017	0.01	-0.06, 0.08	0.0329402
2018	0.03	-0.03, 0.09	0.0390192
2019	0.04	-0.02, 0.10	0.0059950
2021	0.04	-0.03, 0.10	0.0208036
2022	0.03	-0.03, 0.09	0.0429696
<i>Statistic</i>	6.95		—
<i>R²</i>	0.014		—
<i>Adjusted R²</i>	0.012		—
<i>No. Obs.</i>	4,444		—
<i>df</i>	9		—
<i>AIC</i>	7,457		—
<i>Residual df</i>	4,434		—
<i>Sigma</i>	0.559		—

¹CI = Confidence Interval

Table S13. An OLS regression predicting the age gap (Male Age – Female Age) in Google Images of occupation as a function of the yearly pay gap (Male – Female earnings) in occupations according to the U.S. Census (2015 to 2022). The pay gap is not logged given that it is signed (negative values indicate women earn more; positive, men earn more). This model

¹ All Census data related to earnings was downloaded from the following link: <https://www.census.gov/data/tables/time-series/demo/industry-occupation/median-earnings.html>

controls for the logged error of the median yearly earnings (provided by the Census) for both (i) men and (ii) women separately, and (iii) the year of the Census data examined (from 2015 to 2022; note: the 2020 COVID year is missing in the Census data provided). Standard errors are clustered at the occupation level.

We observe a similar statistical robustness in the correlation between the gender pay gap for occupations and the age gap between men and women in Google Images of these occupations. Table S13 presents an OLS regression predicting the age gap (Male Age – Female Age) in Google Images of occupation as a function of the yearly pay gap (Male – Female earnings) in occupations according to the U.S. Census, while controlling for the logged error of the median yearly earnings (provided by the Census) for both (i) men and (ii) women separately, and (iii) the year of the Census data examined. Table S13 shows that, subject to all of these controls, the yearly gender pay gap of an occupation is significantly and positively associated with the age gap depicting men as older than women in Google images of these occupations ($\beta[\text{Gender Pay Gap}] \approx 0.01$, $\text{CI} = [0.00, 0.00]$, $p = 0.02$). While the coefficient is notably small, we find it compelling that there is a robust positive correlation between the ground truth gender pay gap and the age gap we observe in Google Images, given the relatively small sample size of Google Images examined per category, along with the qualitatively distinct natures of these datasets and their corresponding measures.

C. 1. Text Analyses

C. 1. 1 Robustness of the Geometry of Culture Method at the Sentence-level

In prior work focusing on single-word poles and static embeddings, we showed the robustness of the geometry of culture method to reducing or extending the number of single gender terms used to create the respective gender poles. Here, we show that the geometry of culture method is robust to using sentence-level nodes in the creation of both the gender and age poles in contextualized word embedding models. We focus on creating a reduced and extended version of the prompts used to specify the age and gender poles. We focus on examining the results from GPT2-Large, which forms the basis of our main results concerning textual data. To create the reduced form of the age and gender dimension, we simply removed sentences from each pole as defined by our main approach. In the case of the gender dimension, this involved reducing the number of sentences in each pole from 53 to 20; and in the case of the age dimension, this involved reducing the number of sentences in each pole from 22 to 8. To create the extended form of the age and gender dimension, we simply added sentences to each pole as defined by our main approach. In the case of the gender dimension, this involved increasing the number of sentences in each pole to 112; and in the case of the age dimension, this involved increasing the number of sentences in each pole to 44. At the end of this section, we provide the exact sentences used to define the main, extended, and reduced forms of both the gender and age poles.

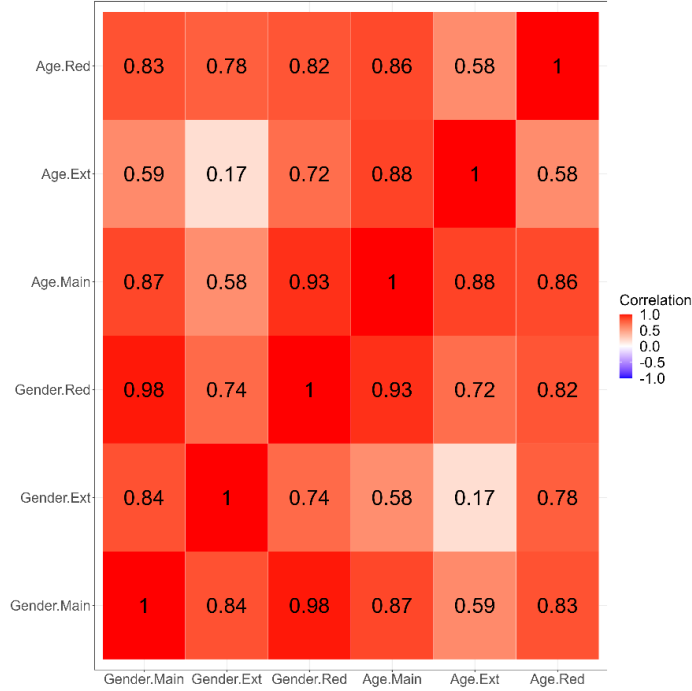


Figure S13. Heatmap displaying the pairwise correlations at the category-level ($n = 3,059$ categories) between and within the main, extended, and reduced constructions of the age and gender dimensions in GPT2-Large. The exact Pearson correlation is indicated within each cell and is calculated using a two-tailed test. All correlations displayed are statistically significant at the $p < 0.0000001$ level. “Main” = pole specification in the main analysis; “Ext” = extended pole specification; “Red” = reduced pole specification.

Figure S13 shows all of the pairwise correlations between and within the main, extended, and reduced constructions of the age and gender dimensions in GPT2-Large. The results show that the main, reduced, and extended versions of the gender dimension are all highly and positive correlated with each other at the $p < 0.0000001$ level; and similarly, the main, reduced, and extended versions of the age dimension are all highly and positive correlated with each other at the $p < 0.0000001$ level. Crucially, this robustness holds when examining correlations across the gender and age dimensions. For all pairwise combination of the main, extended, and reduced form of both the gender and age dimension, we identify a strong, positive, and highly significant correlation between the extent to which a category is associated with men and older ages (all pairwise correlations are significant at the $p < 0.0000001$ level using a Student t-test, two-tailed). This robustness shows that our main results examining GPT2-Large are robust to varying the specification of the sentence-based age and gender poles via the “Geometry of Culture Method” as applied to GPT’s contextualized embeddings.

Here, we provide all of the exact sentences used to define the main, reduced, and extended forms of both the gender and age sentence-level poles in the Geometry of Culture method. Each table also includes the “pole” column which indicates the pole specification to which each sentence belongs: R = Reduced, M = Main, and E = Extended. This same logic is applied to both the gender (Table S14) and the age (Table S15) sentence constructions.

This thing is much more like a man.	This thing is much more like a women.	E
This thing is much more like a boy.	This thing is much more like a girl.	E
This thing is much more like a man in its qualities.	This thing is much more like a women in its qualities.	E
This thing is much more like a boy in its qualities	This thing is much more like a girl in its qualities	E
This thing is much more like a man in its features.	This thing is much more like a women in its features.	E
This thing is much more like a boy in its features	This thing is much more like a girl in its features.	E
This thing is much more like a man in its aspects.	This thing is much more like a women in its aspects.	E
This thing is much more like a boy in its aspects.	This thing is much more like a girl in its aspects.	E
This thing is much more like a man in its characteristics.	This thing is much more like a women in its characteristics.	E
This thing is much more like a boy in its characteristics.	This thing is much more like a girl in its characteristics.	E
This person is a boy.	This person is a girl.	E
This person is a man.	This person is a woman.	E
This person is a male.	This person is a female.	E
This person is a father.	This person is a mother.	E
This person is a son.	This person is a daughter.	E
This person is a brother.	This person is a sister.	E
This person is an uncle.	This person is an aunt.	E
This person is a grandpa.	This person is a grandma.	E
This person is a grandfather.	This person is a grandmother.	E
This person is very masculine.	This person is very feminine.	E
This person has very male qualities.	This person has very female qualities.	E
This person has male energy.	This person has female energy.	E
This person has very male characteristics.	This person has very female characteristics.	E
This person has very male features.	This person has very female features.	E
This person has very masculine qualities.	This person has very feminine qualities.	E
This person has masculine energy.	This person has feminine energy.	E
This person has very masculine characteristics.	This person has very feminine characteristics.	E
This action is manly.	This action is womanly.	E
This behavior is manly.	This behavior is womanly.	E
This attitude is manly.	This attitude is womanly.	E
This feeling is manly.	This feeling is womanly.	E
This idea is manly.	This idea is womanly.	E
This way of being is manly.	This way of being is womanly.	E
This occupation is manly.	This occupation is womanly.	E
This job is manly.	This job is womanly.	E
This responsibility is manly.	This responsibility is womanly.	E
This role is manly.	This role is womanly.	E
This position is manly.	This position is womanly.	E
This task is manly.	This task is womanly.	E
This action is masculine.	This action is feminine.	E
This behavior is masculine.	This behavior is feminine.	E
This attitude is masculine.	This attitude is feminine.	E
This feeling is masculine.	This feeling is feminine.	E
This idea is masculine.	This idea is feminine.	E
This way of being is masculine.	This way of being is feminine.	E
This occupation is masculine.	This occupation is feminine.	E
This job is masculine.	This job is feminine.	E
This responsibility is masculine.	This responsibility is feminine.	E
This role is masculine.	This role is feminine.	E
This position is masculine.	This position is feminine.	E
This task is masculine.	This task is feminine.	E

Table S14. The sentences used to create the gender pole in contextualized embedding models (specifically GPT2-Large). The “pole” column indicates which sentences are included in the reduced (R), main (M), and E (extended) pole specifications.

Young Phrases	Old Phrases	Pole
The person is young.	The person is old.	R,M,E
The person's age is very low.	The person's age is very high.	R,M,E
The person is a baby.	The person is elderly.	R,M,E
The person is youthful.	The person is not youthful.	R,M,E
The person is very youthful.	The person is not very elderly.	R,M,E

The person is a junior.	The person is a senior.	R,M,E
The person is a child.	The person is an adult.	R,M,E
The person is an adolescent.	The person is middle-aged.	R,M,E
The person is very young.	The person is very old.	M,E
The person was born only recently.	The person was born a long time ago.	M,E
He is a boy.	He is a man.	M,E
She is a girl.	She is a woman.	M,E
He is a son.	He is a father.	M,E
She is a daughter.	She is a mother.	M,E
He is a newborn.	He is an adult man.	M,E
She is a newborn.	She is an adult woman.	M,E
He is a young man.	He is an old man.	M,E
She is a young woman.	She is an old woman.	M,E
He is a child.	He is an adult.	M,E
She is a child.	She is an adult.	M,E
He is an adolescent.	He is a middle-aged man.	M,E
She is an adolescent.	She is a middle-aged woman.	M,E
This thing is young.	This thing is old.	E
This thing is very young.	This thing is very old.	E
This style is young.	This style is old.	E
This style is very young.	This style is very old.	E
This style is young.	This style is old.	E
This style is very young.	This style is very old.	E
This thing is for younger people.	This thing is for older people.	E
This feeling is for younger people.	This feeling is for older people.	E
The age of this person is low.	The age of this person is high.	E
The age of this person is lower than others.	The age of this person is higher than others.	E
The age of this person is young.	The age of this person is old.	E
The age of this person is younger.	The age of this person is older.	E
The demeanor of this person is young.	The age of this demeanor is old.	E
The demeanor of this person is younger.	The age of this demeanor is older.	E
This person is the youngest in the room.	This person is the oldest in the room.	E
This person is the most junior in the room.	This person is the most senior in the room.	E
This person has the lowest seniority.	This person has the highest seniority.	E
This person has only been here for a few years.	This person has only been here for many years.	E
This person is junior.	This person is senior.	E
This person is very junior.	This person is very senior.	E
This person is not old enough.	This person is old enough.	E
This person does not satisfy the age requirements.	This person does satisfy the age requirements.	E

Table S15. The sentences used to create the age pole in contextualized embedding models (specifically GPT2-Large). The “pole” column indicates which sentences are included in the reduced (R), main (M), and E (extended) pole specifications.

C. 1. 2 Robustness of GPT2-Large Results Controlling for Linguistic Features of Categories

Here, we show that the strong positive correlation between GPT2-Large’s age and gender associations for social categories, as shown in figure 3 in the main text, is highly robust to controlling for linguistic features of the categories themselves.

Variables	Beta	95% CI ¹	p-value
GPT2-Large Gender Score	0.34	0.34, 0.35	0.000000e+00
Gendered Connotation (0 or 1)	0.00	0.00, 0.00	7.118862e-01
Age Connotation (0 or 1)	0.00	0.00, 0.00	4.325772e-02
Polysemy	0.00	0.00, 0.00	7.966828e-30
Word Frequency (Log Scaled)	0.00	0.00, 0.00	5.712460e-13
<i>Statistic</i>	2,355		
<i>R²</i>	0.772		

Variables	Beta	95% CI ¹	p-value
<i>Adjusted R²</i>	0.771		
<i>No. Obs.</i>	3,488		
<i>df</i>	5		
<i>AIC</i>	-22,640		
<i>Residual df</i>	3,482		
<i>Sigma</i>	0.009		

¹CI = Confidence Interval

Table S16. An OLS regression predicting the age association (Young, -1, to Old, 1) in GPT2-Large for social categories as a function of the gender association (Female, -1, to Male, 1) in GPT2-Large for these same categories, while controlling the linguistic features of these categories, including categories' gender connotation, age connotation, polysemy, and frequency of word usage (log scaled).

Table S16 shows that GPT2-Large's gender associations (-1, female, to 1, male) are highly correlated with GPT2-Large's age associations (-1, young, to 1, old) ($\beta = 0.34$, CI = [0.34, 0.35], $p = 2.2 \times 10^{-16}$), even while controlling for categories' gender connotation, age connotation, polysemy, and word frequency in everyday communication. Importantly, we find that whether a category possesses a gender connotation (e.g. *aunt* connotes a female gender) has no significant correlation with associated age ratings ($\beta = 0.00$, CI = [0.00, 0.00], $p = 0.84$). Whether a category possesses an age connotation has an exceedingly weak association with its age ratings ($\beta = 0.00$, CI = [0.00, 0.00], $p = 0.04$). Polysemy is significantly, though very weakly, correlated with age associations ($\beta = 0.001$, CI = [0.00, 0.00], $p = 2.2 \times 10^{-16}$); and word frequency is significantly, though very weakly, positively correlated with age associations ($\beta = 0.001$, CI = [0.00, 0.00], $p = 1.1 \times 10^{-14}$).

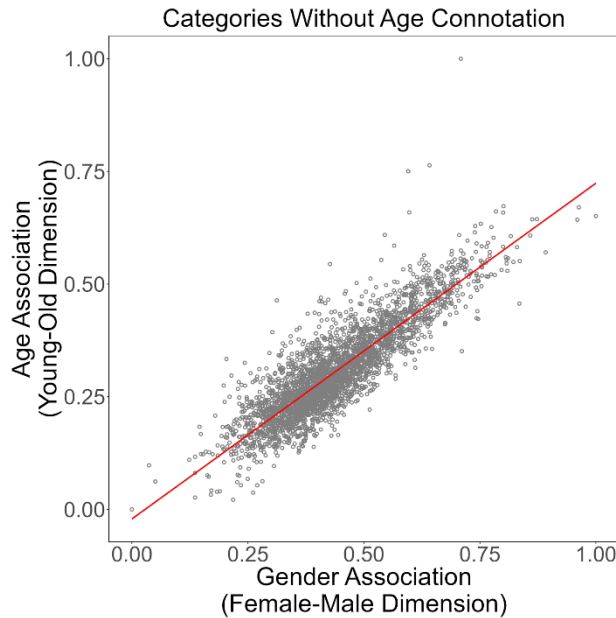


Figure S14. The correlation between age and gender associations in GPT2-Large for all social

categories in Wordnet, while excluding categories that connote age, yielding $n = 3,034$ categories. The horizontal axis presents the gender association from 0 (female) to 1 (male), and the vertical axis presents the age association from 0 (young) to 1 (old). The trend line shows the linear prediction according to an ordinary least squares regression.

We conclude this section by showing that our main analyses of GPT2-Large are nearly identical when excluding all categories that explicitly connote age (e.g., ‘child’ and ‘adult’), yielding 3,034 categories (Figure S14). We would like to note that we took a conservative approach by excluding all categories that moderately implied age (e.g., ‘beginner’ and ‘student’), even though such associations do not always indicate age (older people can still be beginners as a function of context). Nevertheless, even with this conservative approach, we observe that the main correlation reported between the age and gender dimension in GPT2-Large remains strongly and significantly positive, with more male categories being associated with older ages ($r = 0.87$, $t = 100.09$, $p = 2.2 \times 10^{-16}$, Pearson correlation, two-tailed). In fact, the Pearson correlation remains exactly the same.

C. 1. 3 Replication across Nine Language Models

Here we examine age-related gender bias across nine language models, each trained on massive datasets of online text collected from various sources and during distinct periods of time. Specifically, we show that these results replicate when examining a wide range of models, including word2vec¹⁵, GloVe¹⁶, BERT¹⁷, FastText¹⁸, RoBERTa¹⁹, and GPT4²⁰, all of which vary in their dimensionality and data sources, as well as the year in which their training data was collected, ranging from 2013 to 2023. Specifically, we replicate our analyses looking at (i) the canonical word2vec model trained on the Google News corpus in 2013, (ii) a retrained word2vec model on a more recent sample of online news from 2021 to 2023, (iii) the canonical glove model trained on Twitter in 2014, as well as (iv) on Wikipedia in 2015, (v) Meta’s 2016 FastText model trained on Wikipedia and the common crawl, (vi) Google’s 2018 BERT model trained on Wikipedia and digitized books, (vii) the 2019 RoBERTa (Robustly Optimized BERT approach) model trained on Wikipedia, digitized books, online news, Reddit, and the common crawl, and finally, (viii) OpenAI’s 2023 GPT4 model trained on Wikipedia, the common crawl, digitized books, and Reddit (specifically, we used OpenAI’s most updated state-of-the-art embeddings as of early 2024, which is commonly described as the embeddings for GPT4, but OpenAI is not transparent about the design of GPT4, so our analyses are best described as using OpenAI’s most state-of-the-art embeddings to date). Combined, these language models capture a vast sample of the internet, both across platforms and across time, such that the evidence of age-related gender bias across all these models is a powerful testament to its striking cultural pervasiveness.

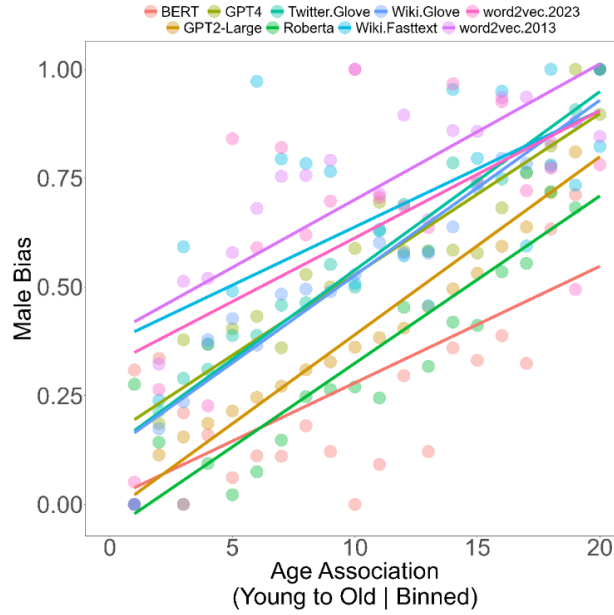


Figure S15. Women are represented as significantly younger than men in billions of words encoded by nine different language models, trained on data from online news, Wikipedia, Twitter, and a random scrape of the web. The correlation between young-old age associations and male bias gender associations at the category-level across popular wording embedding models. The horizontal axis presents age associations as binned into 20 evenly spaced bins (calculated separately for each word embedding model). The vertical axis displays male bias as the min-max normalized representation of the female-male dimension (min-max normalized is applied across the mean values for each bin). Trend lines show the linear prediction according to an ordinary least squares regression calculated over the data associated with each word embedding model separately. Data points show the mean value of male bias for each model calculated within each age bin.

As Figure S15 shows, in all models we observe a strong positive correlation between the associated age of each category and the extent to which it is associated with men. Social categories associated with higher ages are more likely to be male according to the following models: the canonical word2vec model trained on the Google News corpus in 2013 (raw correlation: $r = 0.17$, $t = 9.67$, $CI = [0.13-0.21]$, $p = 2.2 \times 10^{-16}$, $n = 3056$ categories; correlation between age bins and average gender score: $r = 0.77$, $t = 5.15$, $CI = [0.50-0.90]$, $p = 6.9 \times 10^{-6}$, $n = 20$ bins), a retrained word2vec model on a more recent sample of online news from 2021 to 2023 (raw correlation: $r = 0.10$, $t = 5.55$, $CI = [0.06-0.14]$, $p = 2.2 \times 10^{-8}$, $n = 3049$ categories; correlation between age bins and average gender score: $r = 0.60$, $t = 3.2$, $CI = [0.22-0.82]$, $p = 0.004$, $n = 20$ bins), the canonical glove model trained on Twitter in 2014 (raw correlation: $r = 0.35$, $t = 19.75$, $CI = [0.32-0.39]$, $p = 2.2 \times 10^{-16}$, $n = 2714$ categories; correlation between age bins and average gender score: $r = 0.96$, $t = 16.3$, $CI = [0.92-0.99]$, $p = 3.1 \times 10^{-12}$, $n = 20$ bins), as well as on Wikipedia in 2015 (raw correlation: $r = 0.38$, $t = 18.23$, $CI = [0.34-0.41]$, $p = 2.2 \times 10^{-16}$, $n = 3111$ categories; correlation between age bins and average gender score: $r = 0.96$, $t = 15.89$, $CI = [0.91-0.99]$, $p = 4.87 \times 10^{-12}$, $n = 20$ bins), Meta's 2016 FastText model trained on Wikipedia and the common crawl (raw correlation: $r = 0.18$, $t = 5.7$, $CI = [0.06-0.14]$, $p = 1.24 \times 10^{-8}$, $n = 3220$ categories; correlation between age bins and average gender score: $r = 0.61$, $t = 3.25$, $CI = [0.23-0.83]$, $p = 0.004$, $n = 20$ bins), Google's 2018 BERT model trained on Wikipedia and

digitized books (raw correlation: $r = 0.14$, $t = 8.3$, $CI = [0.11-0.17]$, $p = 2.2 \times 10^{-16}$, $n = 3410$ categories; correlation between age bins and average gender score: $r = 0.64$, $t = 3.5$, $CI = [0.28-0.85]$, $p = 0.002$, $n = 20$ bins), the 2019 RoBERTa (Robustly Optimized BERT approach) model trained on Wikipedia, digitized books, online news, Reddit, and the common crawl (raw correlation: $r = 0.2$, $t = 10.64$, $CI = [0.16-0.23]$, $p = 2.2 \times 10^{-16}$, $n = 2751$ categories; correlation between age bins and average gender score: $r = 0.88$, $t = 7.99$, $CI = [0.72-0.95]$, $p = 2.45 \times 10^{-7}$, $n = 20$ bins), OpenAI’s open source GPT2-large model trained on Wikipedia, the common crawl, digitized books, and Reddit (raw correlation: $r = 0.88$, $t = 7.99$, $CI = [0.72-0.95]$, $p = 2.45 \times 10^{-7}$, $n = 3495$ categories; correlation between age bins and average gender score: $r = 0.97$, $t = 16.46$, $CI = [0.92-0.98]$, $p = 2.69 \times 10^{-12}$, $n = 20$ bins); and finally, OpenAI’s closed-source GPT4 model trained on Wikipedia, the common crawl, digitized books, and Reddit (raw correlation: $r = 0.17$, $t = 10.25$, $CI = [0.14-0.20]$, $p = 2.2 \times 10^{-16}$, $n = 3495$ categories; correlation between age bins and average gender score: $r = 0.93$, $t = 10.57$, $CI = [0.82-0.97]$, $p = 3.72 \times 10^{-9}$, $n = 20$ bins), (all tests reported use Pearson correlation, two-tailed).

C. 1. 4 Robustness of the Age Dimension across Models

In a recent publication using the same Google and Wikipedia datasets¹⁴, we showed that the gender dimension results across all models examined in the current study are robust to a variety of ways of specifying the gender dimension, including reducing the gender dimension to only three words per pole and increasing it to 15 words per pole (see Fig. S13 of Guilbeault et al. 2024¹⁴). In this prior paper, we also showed that the gender associations for social categories are highly correlated across a wide range of models. If we rerun the same analyses looking across all the models used in the same study, we continue to find that models’ gender associations are highly correlated with $r = 0.38$ on average at the $p < 0.001$ level (Pearson correlation, two-tailed). Here, we run these same analyses to evaluate the robustness of the age dimension across models.

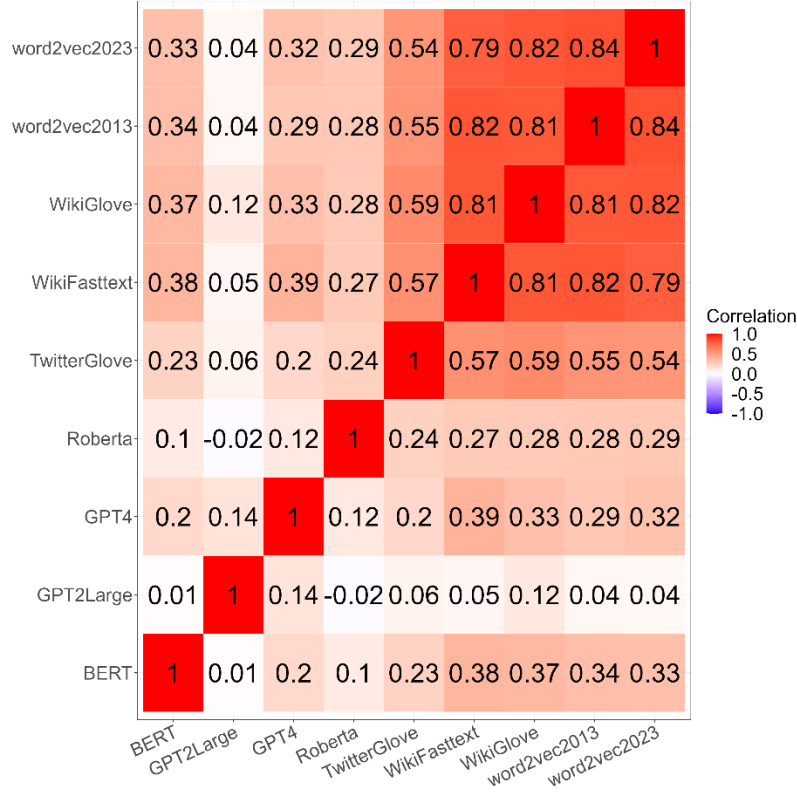


Figure S16. Heatmap displaying the pairwise correlations in age associations (paired at the level of social categories) across all models examined.

Figure S16 presents a heatmap displaying the pairwise correlations in age associations (paired at the level of social categories) across all models examined. The results show that nearly all models exhibit significant positive correlations in how the age dimension within each model associates age with the same set of social categories. The average correlation in age association across models (paired at the social category level) is $r = 0.42$ and is significant at the $p < 0.001$ level (Pearson correlation, two-tailed). These results suggest that the age dimension captures common age-based stereotypical associations across the different datasets and training techniques employed by these different models. It is important to note that very high correlations are not necessarily expected between models, given that they were trained on quite distinct datasets (e.g., Twitter vs. Wikipedia vs. Google News) and during different time periods. This variation makes it all the more notable that the significant positive correlation between gender and age replicates across all of these models, and that the age associations detected by the geometry of culture methods exhibit comparable levels of correlation across models as exhibited by the gender dimension¹⁴.

C. 1. 5 Benchmarking GPT2-Large Demographic Dimensions Against the Census

Here, we test whether social categories' association along the age and gender dimension in GPT2-Large predicts the ground truth median age estimates of occupations according to the U.S. Census. We were able to match 867 categories from the GPT2-large embedding dataset to occupational categories in the U.S. Census, and we focus on comparing against census data from a relevant time frame, specifically from 2020 to 2023 (where the 2023 Census record is the most recent to date). We begin by using an OLS regression to evaluate the association between an

occupation's position along the age dimension in GPT2-Large's embedding space and the median age of people in this occupation according to the Census. The results of this model are presented in table S14, below. Note, the correlation between the position of occupational categories along the age dimension in GPT2-Large's embedding space and the median age of people in organizations from a nationally representative sample is *not* expected to be particularly high, given the many foundational differences in how age is measured and sampled across these different data structures. In this case, we are interested in observing whether there is any reasonable and significant correlation in the expected direction, where even a moderately strong correlation is considered impressive under this conservative analysis.

Variables	Beta	95% CI ¹	p-value
GPT2-Large Age Association	21.24	3.37, 39.11	0.01985
Census year			—
2020	—	—	—
2021	-0.18	-0.34, -0.02	0.02166
2022	0.11	-0.05, 0.27	0.20116
2023	-0.27	-0.48, -0.06	0.01103
<i>Statistic</i>	5.59		—
<i>R</i> ²	0.008		—
<i>Adjusted R</i> ²	0.006		—
<i>No. Obs.</i>	2,845		—
<i>df</i>	4		—
<i>AIC</i>	16,986		—
<i>Residual df</i>	2,840		—
<i>Sigma</i>	4.78		—

¹CI = Confidence Interval

Table S17. An OLS regression predicting the median age of people in occupations according to the U.S. Census (2020 to 2023) as a function of the age association (Young, -1, to Old, 1) in GPT2-Large for these same occupations, while controlling for Census year and clustering standard errors at the occupation category level.

Table S17 displays an OLS regression predicting the median age of people in occupations according to the U.S. Census (2020 to 2023) as a function of the age association (Young, -1, to Old, 1) in GPT2-Large for these same occupations, while controlling for Census year and clustering standard errors at the occupation category level. The results of table S17 show that the age association (-1, young, to 1, old) of occupations in GPT2-Large's embedding space are positively correlated with the median age of people in these occupations according to the U.S. Census, across years 2020 to 2023 ($\beta = 21.24$ years, $CI = [3.37, 39.11]$, $p = 0.02$). For reference, the raw correlation between the GPT2-Large's age score for a given occupation and this occupation's median age in the census across years is $r = 0.08$ ($p = 0.0001$, $t = 4.42$, Pearson correlation, two-tailed).

Variables	Beta	95% CI ¹	p-value
GPT2-Large Gender Association	6.4	2.8, 10	6.151746e-04
Census Year			—
2020	—	—	—
2021	-0.18	-0.69, 0.32	4.701926e-01
2022	0.11	-0.39, 0.61	6.793184e-01
2023	-0.27	-0.77, 0.22	2.794247e-01
<i>Statistic</i>	3.64		—
<i>R²</i>	0.005		—
<i>Adjusted R²</i>	0.004		—
<i>No. Obs.</i>	2,845		—
<i>df</i>	4		—
<i>AIC</i>	16,994		—
<i>Residual df</i>	2,840		—
<i>Sigma</i>	4.79		—

¹CI = Confidence Interval

Table S18. An OLS regression predicting the median age of people in occupations according to the U.S. Census (2020 to 2023) as a function of the gender association (Female, -1, to Male, 1) in GPT2-Large’s embedding space for these same occupations, while controlling for Census year.

We next examine whether the gender association for an occupation in GPT2-Large’s embedding space is significantly associated with the median age of people in this occupation according to the U.S. Census. Table S18 displays the results of an OLS regression predicting the median age of people in occupations according to the U.S. Census (2020 to 2023) as a function of the gender association (Female, -1, to Male, 1) in GPT2-Large for these same occupations, while controlling for Census year. The results of table S15 show that the gender of occupations in GPT2-Large’s embedding space is positively correlated with the median age of people in these occupations according to the U.S. Census, across years 2020 to 2023 ($\beta = 6.4$ years, CI = [2.8, 10], $p = 0.0004$). That this shows a positive correlation indicates that occupations more associated with men in GPT2-Large are also significantly associated with higher median ages in the U.S. Census. For reference, the raw correlation between the GPT2-Large’s gender score for a given occupation and this occupation’s median age in the census across years is $r = 0.08$ ($p = 0.0006$, $t = 3.42$, Pearson correlation, two-tailed). However, this correlation is not particularly robust when clustering standard errors at the occupation level. Replicating Table S18 while clustering standard errors at the occupation level reduces the significance of the main correlation to just outside of conventional levels of statistical significance,

though the direction and magnitude of the slope remains stable ($\beta = 6.4$ years, $CI = [-0.64, 13.53]$, $p = 0.07$).

D. 1 Experimental Robustness

D. 1. 1 Robustness to Controlling for Participants' Gender and Age

Here, we show that our experimental results are robust to controlling for the self-identified gender and age of participants. Table S19 presents an OLS model that predicts participants' age estimates for each occupation as a function of whether they were randomized to the control of image condition, if in the image condition, whether they uploaded a Female or Male face, with fixed effects by occupation, participants' gender, and participants' age. We find that, even with these controls, uploading a female face continues to be strongly associated with providing a lower age estimate for the occupation depicted ($\beta[\text{Female}] = -1.8$, $CI = [-2.2, -1.4]$, $p = 2.62 \times 10^{-19}$), while uploading a male face continues to be strongly associated with providing a higher age estimate for the occupation depicted ($\beta[\text{Male}] = 0.75$, $CI = [0.33, 1.2]$, $p = 5.45 \times 10^{-4}$). Importantly, there is no statistically significant relationship between participants' gender ($\beta[\text{Male}] = -0.04$, $CI = [-0.37, 0.28]$, $p = 0.78$) or age ($\beta = -0.01$, $CI = [-0.02, 0.0]$, $p = 0.07$) in terms of predicting their average age estimates for each occupation.

Variables	Beta	95% CI ¹	p-value
Gender of Image			
Control	—	—	—
Female	-1.8	-2.2, -1.4	2.626077e-19
Male	0.75	0.33, 1.2	5.457407e-04
Occupation			
aeronautical engineer	—	—	—
applied scientist	-0.91	-2.6, 0.76	2.858982e-01
art student	-18	-19, -16	1.130295e-99
art teacher	-3.9	-5.5, -2.2	4.617140e-06
astronaut	-0.21	-1.9, 1.5	8.030376e-01
astrophysicist	5.4	3.7, 7.0	2.500162e-10
automotive engineer	-3.5	-5.2, -1.8	5.141495e-05
bookkeeper	-0.86	-2.5, 0.75	2.949562e-01
cardiologist	6.7	5.0, 8.3	2.159901e-15
chief executive officer	10	8.5, 12	6.013130e-33
chiropractor	-0.21	-1.9, 1.4	8.020545e-01
clarinetist	-4.7	-6.3, -3.1	5.885215e-09
climatologist	-0.34	-2.0, 1.3	6.908522e-01
computer expert	-6.0	-7.6, -4.3	6.674720e-13
cosmetic surgeon	4.3	2.6, 5.9	5.373939e-07
dietician	-3.9	-5.5, -2.2	3.669508e-06
dressmaker	-1.1	-2.8, 0.48	1.679529e-01

Variables	Beta	95% CI¹	p-value
editor in chief	6.8	5.1, 8.4	5.947662e-16
educator	-3.4	-5.1, -1.8	4.285821e-05
english professor	7.2	5.6, 8.9	2.251872e-17
english teacher	-3.1	-4.7, -1.5	1.518845e-04
fashion designer	-4.9	-6.5, -3.2	6.794577e-09
fashion model	-15	-17, -13	2.503759e-64
financial analyst	-1.1	-2.7, 0.56	1.994036e-01
geneticist	0.69	-0.98, 2.4	4.202374e-01
graphic designer	-8.9	-11, -7.3	2.659920e-26
harpist	-3.7	-5.4, -2.0	2.432671e-05
hygienist	-6.7	-8.4, -5.1	2.867019e-15
immunologist	2.5	0.84, 4.2	3.201308e-03
intellectual	4.2	2.6, 5.9	3.056443e-07
intelligence agent	-0.56	-2.2, 1.1	5.020653e-01
intelligence analyst	-0.62	-2.3, 1.1	4.664191e-01
interior decorator	-4.8	-6.4, -3.1	9.120314e-09
literary agent	3.0	1.3, 4.6	3.687730e-04
logician	3.7	2.1, 5.3	8.200937e-06
marine engineer	-0.41	-2.1, 1.3	6.355475e-01
mathematician	5.1	3.3, 6.8	9.726290e-09
media consultant	-6.2	-7.8, -4.6	1.136291e-13
model	-15	-17, -14	4.783899e-75
neurobiologist	2.9	1.2, 4.5	5.745750e-04
number theorist	2.9	1.2, 4.5	6.985623e-04
nurse practitioner	-4.3	-5.9, -2.6	3.530912e-07
painter	-5.6	-7.2, -3.9	1.935463e-11
pianoplayer	-4.7	-6.4, -3.0	3.241741e-08
poet	-0.01	-1.6, 1.6	9.870025e-01
professional dancer	-12	-14, -11	2.976854e-49
professor	7.1	5.5, 8.7	8.301593e-18
programmer	-8.6	-10, -7.0	1.991510e-25
school teacher	-5.2	-6.8, -3.5	8.594563e-10
screen actor	-6.0	-7.7, -4.4	2.796013e-13
singer	-10	-12, -8.7	7.226814e-34
social worker	-4.1	-5.7, -2.5	5.353472e-07
systems analyst	-5.6	-7.2, -3.9	8.763917e-11
trained nurse	-5.0	-6.7, -3.4	3.152731e-09
Participant Age	-0.01	-0.02, 0.00	7.520983e-02

Variables	Beta	95% CI ¹	p-value
Participant Gender			
Female	—	—	—
Male	-0.04	-0.37, 0.28	7.881101e-01
<i>Statistic</i>	95.6		
<i>R²</i>	0.354		
<i>Adjusted R²</i>	0.350		
<i>No. Obs.</i>	9,999		
<i>df</i>	57		
<i>AIC</i>	70,504		
<i>Residual df</i>	9,941		
<i>Sigma</i>	8.20		

¹CI = Confidence Interval

Table S19. An OLS regression predicting participants' average age estimates of occupations as a function of their own gender and age, with fixed effects by occupation.

As an additional robustness test, we examine whether participants' image choices and age estimated are biased by whether the images they upload match their own gender, as well as the distance between their own age and the age of the occupation estimated. This analysis is restricted only to subjects in the image condition because only participants in this condition uploaded images depicting people in occupations with related demographic attributes. Table S20 presents an OLS model that predicts participants' age estimates for each occupation as a function of whether they uploaded a Female or Male face, whether the gender of this face matches their own, the distance between their own age and the age of the occupation estimated, while including fixed effects by occupation. We find that, subject to these controls, uploading a male face continues to be strongly associated with providing a higher age estimate for the occupation depicted compared to uploading a female face ($\beta[\text{Male}] = 2.9$, $\text{CI} = [2.4, 3.5]$, $p = 1.92 \times 10^{-23}$). Importantly, whether participants' own gender matches the gender of their uploaded face has no significant relationship with the age estimate they provided ($\beta[\text{TRUE Match}] = -0.15$, $\text{CI} = [-0.61, 0.31]$, $p = 0.51$). Similarly, the absolute distance between participants' own age and the age estimate they provided for each occupation carried no significant relationship with the age estimated provided ($\beta = 0.00$, $\text{CI} = [-0.02, 0.02]$, $p = 0.69$).

Variables	Beta	95% CI ¹	p-value
Gender of Image			
Female	—	—	—
Male	2.9	2.4, 3.5	1.923759e-23
Occupation			
applied scientist	-0.70	-3.0, 1.6	5.530478e-01
art student	-16	-18, -14	1.008206e-41
art teacher	-0.49	-2.8, 1.8	6.753811e-01

Variables	Beta	95% CI¹	p-value
astronaut	2.0	-0.32, 4.4	9.028783e-02
astrophysicist	8.0	5.7, 10	1.432618e-11
automotive engineer	-3.0	-5.2, -0.70	1.024657e-02
bookkeeper	-0.45	-2.8, 1.9	7.052424e-01
cardiologist	8.1	5.7, 10	1.625856e-11
chief executive officer	11	9.2, 14	1.819230e-22
chiropractor	0.38	-2.0, 2.7	7.535070e-01
clarinetist	-1.7	-3.9, 0.49	1.267347e-01
climatologist	0.81	-1.5, 3.1	4.862031e-01
computer expert	-6.4	-8.7, -4.1	6.824076e-08
cosmetic surgeon	5.6	3.2, 7.9	3.455917e-06
dietician	-1.2	-3.7, 1.2	3.127914e-01
dressmaker	0.34	-1.9, 2.6	7.708824e-01
editor in chief	8.2	5.9, 10	1.501983e-12
educator	-2.4	-4.8, -0.01	4.862860e-02
english professor	8.8	6.5, 11	2.213340e-13
english teacher	-2.5	-4.7, -0.20	3.312367e-02
fashion designer	-3.4	-5.8, -1.1	3.794814e-03
fashion model	-11	-13, -8.4	1.957093e-18
financial analyst	-0.48	-2.8, 1.8	6.784586e-01
geneticist	1.5	-0.87, 3.8	2.178234e-01
graphic designer	-7.7	-10, -5.4	5.779674e-11
harpist	-1.4	-3.8, 0.94	2.395502e-01
hygienist	-4.4	-6.8, -2.0	2.533398e-04
immunologist	4.8	2.5, 7.2	6.020042e-05
intellectual	6.7	4.4, 8.9	8.007450e-09
intelligence agent	0.24	-2.0, 2.5	8.362571e-01
intelligence analyst	0.27	-2.0, 2.6	8.145163e-01
Interior decorator	-1.9	-4.2, 0.35	9.722221e-02
literary agent	5.1	2.7, 7.4	2.364061e-05
logician	7.8	5.6, 10	5.067816e-12
marine engineer	-0.13	-2.5, 2.2	9.128508e-01
mathematician	10	7.6, 12	1.308002e-16
media consultant	-5.5	-7.8, -3.1	6.691242e-06
model	-12	-14, -9.7	1.156254e-23
neurobiologist	4.6	2.3, 6.9	7.055090e-05
number theorist	4.8	2.5, 7.1	4.262006e-05
nurse practitioner	-1.8	-4.0, 0.49	1.239345e-01

Variables	Beta	95% CI ¹	p-value
painter	-3.6	-5.9, -1.3	2.393759e-03
pianoplayer	-2.7	-5.0, -0.43	1.986196e-02
poet	4.2	1.9, 6.4	2.791559e-04
professional dancer	-10	-13, -7.8	1.807987e-16
professor	9.5	7.3, 12	3.248155e-16
programmer	-7.1	-9.3, -4.9	4.751122e-10
school teacher	-2.9	-5.2, -0.66	1.143949e-02
screen actor	-4.1	-6.3, -1.9	3.062595e-04
singer	-7.8	-10, -5.4	7.020420e-11
social worker	-1.0	-3.3, 1.3	3.923490e-01
systems analyst	-5.3	-7.7, -2.8	2.404048e-05
trained nurse	-2.7	-5.0, -0.35	2.450996e-02
Self-Gender Match Image			
FALSE	—	—	—
TRUE	-0.15	-0.61, 0.31	5.161807e-01
[(Self-Age – Est. Age of Occ.)	0.00	-0.02, 0.02	6.894237e-01
<i>Statistic</i>	51.9		
<i>R</i> ²	0.378		
<i>Adjusted R</i> ²	0.371		
<i>No. Obs.</i>	4,829		
<i>df</i>	56		
<i>AIC</i>	33,956		
<i>Residual df</i>	4,772		
<i>Sigma</i>	8.09		

¹CI = Confidence Interval

Table S20. An OLS regression predicting participants’ average estimate of the age of occupations as a function of whether the image they uploaded matches their own gender and of the similarity between their own age and the age estimate they provided (only participants in the Image condition are included in this analysis). Fixed effects by occupation are included.

D. 1. 2 Robustness to Alternative Experimental Measures of Gender Association

Here, we show that the results presented in panel C of Figure 2 are robust to using alternative experimental measures of gender association, including the behavioral results of a separate experiment with a similar design. In this analysis, we measure the gender association of each occupation using data from two experiments: (i) the experiment presented in the main text, and (ii) a recently published experiment with a nearly similar design except that participants in this published experiment were not asked to estimate the average age of occupations or report their hiring preferences; instead, they simply rated the gender most associated with the same set of occupations, which they did using a manual slider and also via the images they uploaded to the experiment¹⁴.

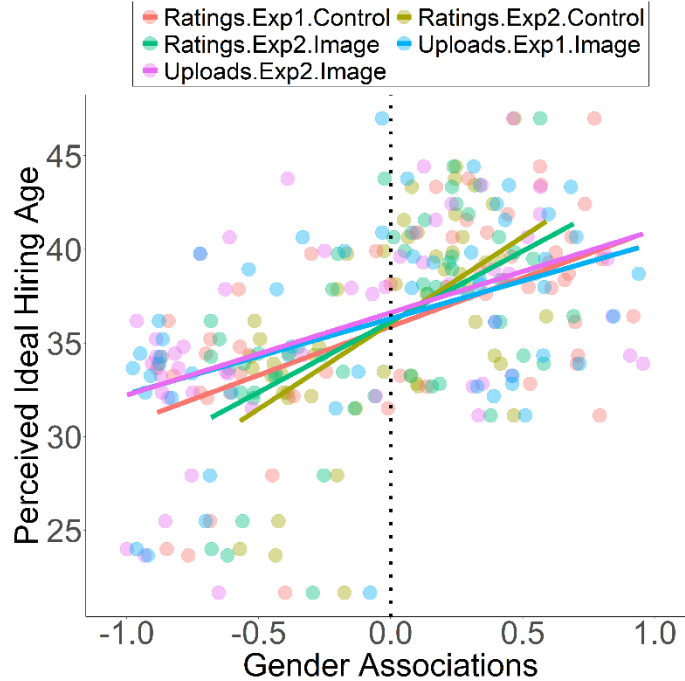


Figure S17. Gender associations predict age-related biases in hiring preferences across occupations (positive values indicate male associations). Participants ($n = 465$) from a nationally representative sample were randomized either to the “Image” condition, in which they googled for images of occupations ($n = 54$), or the “Control” condition, in which they googled for image-based descriptions of random categories (e.g. *apple*) unrelated to occupations. Figure shows the correlation between the gender association and the perceived ideal hiring age of each occupation (averaged across all participants in the control condition); this correlation is shown for measures of the gender association of each occupation according to: participants manual gender ratings in the (i) control and (ii) image condition of the current experiment, in the (iii) control and (iv) image condition of a prior experiment, and in the images uploaded by participants in the current (v) and (vi) prior experiment. Data points show the average gender association and perceived ideal hiring age for each occupation, according to each measure.

Figure S17 shows that across a number of methods and datasets for measuring the gender association of each occupation, occupations that are more female (male) are significantly associated with participants reporting lower (higher) recommended ages for hiring. Using data from our main experiment, we show that control participants’ perceived ideal age for hireability is strongly and positively correlated with the extent to which each occupation is associated with men as measured by (i) control participants’ manual gender ratings of occupations (“Ratings.Exp1.Control” in Fig. S17; $r = 0.58$, $p = 3.52 \times 10^{-6}$), and by (ii) the gender associations in the images uploaded by participants in the Image condition (“Uploads.Exp1.Image” in Fig. S13; $r = 0.45$, $p = 0.0006$), (Pearson correlation, two-tailed). These results are robust to measuring the gender associations of these occupations using the manual ratings and content uploads of participants in a prior experiment¹⁴ with a nearly identical design that involved the same occupations but did not ask participants to report their age estimates. Figure S17 further shows that control participants’ perceived ideal age for hiring is also strongly and

positively correlated with the extent to which each occupation is associated with men according to participants' manual gender ratings of occupations in both (i) the control ("Ratings.Exp2.Control" in Fig. S17; $r = 0.58, p = 4.80 \times 10^{-6}$) and (ii) the image condition ("Ratings.Exp2.Image" in Fig. S13; $r = 0.57, p = 5.79 \times 10^{-6}$) from this prior experiment, as well as with (iii) the images uploaded by the same participants from the image condition ("Uploads.Exp2.Image" in Fig. S17; $r = 0.50, p = 0.0001$) (Pearson correlation, two-tailed, $n = 54$ occupations for each correlation). Our results are thus highly robust both to the datasets and measurement strategy used to quantify the gender associated with each occupation, including participants' manual gender ratings across experiments, as well as the gender biases in the images participants uploaded for occupations across separate experiments.

D. 1. 3 Robustness to Likert Measure of Perceived Hireability

In addition to collecting participants' average age estimates in the treatment condition, we also asked participants to rate the perceived hireability of the image they uploaded along a standard 7-point Likert scale (from "Not at all hireable" to "Extremely hireable"). This analysis only applies to subjects in the treatment (i.e., the "Image") condition.

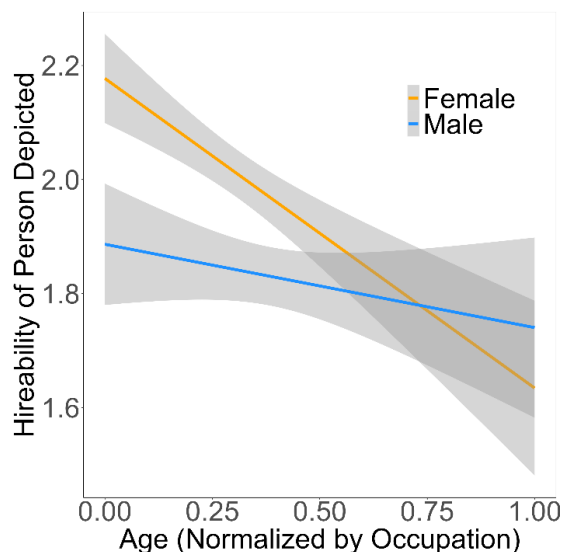


Figure S18. Age-related gender bias in $n = 224$ (Image Condition only) participants' explicit ratings of the hireability of the person represented in the image they uploaded for each occupation. Only participants in the Image condition are examined in this analysis. This figure shows the perceived hireability of the image uploaded for each occupation (indicated by each participant using a 7-point Likert scale), as a function of the gender of this image and its associated age estimate (normalized from 0 to 1 for each occupation). The horizontal axis captures each participant's estimated age for the occupation depicted; participants' estimates are min-max normalized at the occupation level across all participants. Error bands display 95% confidence intervals calculated using a t-test (two-tailed).

Figure S18 shows that for men, there is no correlation between perceived age and hireability ($r_s = 0.00, p = 0.65$, Spearman Correlation, two-tailed). Yet, for women, there is a highly significant and negative correlation between age and perceived hireability ($r_s = -0.07, p = 7.08 \times 10^{-5}$, Spearman Correlation two-tailed): that is, younger women are rated as much more hireable than older women. As the estimated age reaches the upper decile for each occupation, we see an inversion, where older

men are rated as more hireable than older women. An unexpected result in this analysis is that women appear to be rated as more hireable overall, controlling for age estimate; yet importantly, the interaction between age and gender remains consistent with our findings relating to gendered ageism, in the sense that only women are penalized for being older, not men. In general, Likert responses are coarse-grained and vulnerable to social desirability bias (whereby participants provide answers they think the researchers want to receive); but nevertheless, the core patterns are consistent with our main findings, suggesting our results are robust to the style of question used in our survey instrument.

D. 1. 4 Benchmarking Participants' Age and Gender Judgments Against the U.S. Census Data

Here, we test whether experimental participants' judgments of the average age of people in occupations are consistent with ground truth data on the actual age of people in these same occupations according to the U.S. Census, from 2020 to 2023. We focus on participants in the control condition whose judgments were not primed by exposure to images of occupations prior to providing their age judgments (though correlations also hold as expected in the treatment condition; results presented below). Each occupation in the experimental search task could be matched to an occupation in the Census from 2020 to 2023 associated with the median age of people belonging to this occupation.

Variables	Beta	95% CI ¹	p-value
Participant's Age Estimate	0.06	0.05, 0.06	2.309538e-123
Census Year			
2020	—	—	—
2021	0.23	0.11, 0.34	9.261713e-05
2022	0.23	0.12, 0.35	4.742680e-05
2023	-0.14	-0.25, -0.03	1.527270e-02
Subject Fixed Effects			Included
<i>Statistic</i>	5.31		—
<i>R²</i>	0.053		—
<i>Adjusted R²</i>	0.043		—
<i>No. Obs.</i>	22,806		—
<i>df</i>	238		—
<i>AIC</i>	116,221		—
<i>Residual df</i>	22,567		—
<i>Sigma</i>	3.08		—

¹CI = Confidence Interval

Table S21. An OLS regression predicting the median age of people in an occupation according to the U.S. Census (years 2020 to 2023) as a function of experimental participants' estimates of the average age of people in this occupation. Participant fixed effects are included. Only data from control participants are presented here.

Table S21 presents an OLS regression predicting the median age of people in an occupation according to the U.S. Census (years 2020 to 2023) as a function of experimental participants' estimates of the average age of people in this occupation. The model shows that participants' age estimates are significantly correlated with the median age of occupations according to the U.S. Census ($\beta = 0.06$, $CI = [0.05, 0.06]$, $p = 2.309538 \times 10^{-123}$), while controlling for the Census year (2020-2023) and participant. For reference, the raw correlation between participants' age estimates and the medium age of people in the same occupations according to the U.S. census is $r = 0.13$ ($p = 2.2 \times 10^{-16}$, Pearson correlation, two-tailed). Replicating the same OLS in Table S15 but using the age estimates of participants in the treatment condition yields a comparably significant positive correlation between participants' estimates and the median Census age for each occupation ($\beta = 0.045$, $CI = [0.04, 0.05]$, $p = 2.2 \times 10^{-16}$). For reference, the raw correlation between treated participants' age estimates and the medium age of people in the same occupations according to the U.S. census is $r = 0.10$ ($p = 2.2 \times 10^{-16}$, Pearson correlation, two-tailed).

Variables	Beta	95% CI ¹	p-value
Participant's Perceived Ideal Age of Hire	0.05	0.04, 0.05	1.991795e-68
Census Year			
2020	—	—	—
2021	0.22	0.11, 0.34	1.240520e-04
2022	0.23	0.12, 0.34	6.094483e-05
2023	-0.14	-0.25, -0.03	1.525990e-02
<i>Subject Fixed Effects</i>			Included
<i>Statistic</i>	4.19		—
<i>R²</i>	0.042		—
<i>Adjusted R²</i>	0.032		—
<i>No. Obs.</i>	22,806		—
<i>df</i>	238		—
<i>AIC</i>	116,476		—
<i>Residual df</i>	22,567		—
<i>Sigma</i>	3.09		—

¹CI = Confidence Interval

Table S22. An OLS regression predicting the median age of people in an occupation according to the U.S. Census (years 2020 to 2023) as a function of experimental participants' perceptions of the ideal hiring age of people in this occupation. Participant fixed effects are included. Only data from control participants are shown since only control participants provided judgments regarding ideal hiring ages.

Table S22 replicates table S21 while comparing participants' perceptions of the ideal hiring age for people in each occupation and the median age associated with each occupation according to the U.S. census. The model shows that participants' perceptions of the ideal hiring age for

occupations are significantly correlated with the median age of people in these occupations according to the U.S. Census ($\beta = 0.05$, $CI = [0.04, 0.05]$, $p = 1.991795 \times 10^{-68}$), while controlling for the Census year (2020-2023) and participant. For reference, the raw correlation between participants' age estimates and the medium age of people in the same occupations according to the U.S. census is $r = 0.105$ ($p = 2.2 \times 10^{-16}$, Pearson correlation, two-tailed). Participants in the treatment condition are not examined because they did not provide their judgments concerning the ideal hiring age of candidates for particular occupations.

D. 1. 5 Full Summary of Preregistered Hypotheses and Results

In our manuscript, we report all results associated with each of our main hypotheses as preregistered (see <https://osf.io/x9scm>) except for "Main Hypothesis 3," which we deemed to be more complex than needed to make our main argument, especially in light of the number of other analyses and datasets our manuscript compiles. To exhaustively indicate the alignment between our preregistered hypotheses and our results, here we present the results of the preregistered hypothesis associated with Main Hypothesis 3, which stated:

"When participants upload an image of a male for a given occupation, their estimate for the average age of this occupation is going to be significantly closer to the perceived ideal age of this occupation (reported by participants in the control condition) than when participants upload an image of a woman for the same occupation. This includes when controlling for whether participants in the control condition report viewing an occupation as male or female typed."

Table S23 presents precisely the analysis proposed for testing Main Hypothesis 3. Specifically, Table S23 uses an OLS regression to predict the absolute age gap between a participant's age estimate for an occupation in the Image condition and the average ideal age associated with each occupation in the control condition. Furthermore, as proposed, this model controls for whether the majority of subjects in the control condition coded an occupation as male or female, along with fixed effects by participant (in the Image condition) and occupation.

Variables	Beta	95% CI ¹	p-value
Gender of Uploaded Image			
Female	—	—	—
Male	0.72	0.35, 1.1	1.340103e-04
Male Occupation (Control Ratings)			
FALSE	—	—	—
TRUE	-0.28	-1.8, 1.2	7.120094e-01
Participant Fixed Effects			Included
Occupation Fixed Effects			Included
<i>Statistic</i>	7.50		
<i>R</i> ²	0.313		

Variables	Beta	95% CI ¹	p-value
Adjusted R ²	0.272		
No. Obs.	4,829		
df	277		
AIC	29,533		
Residual df	4,551		
Sigma	5.01		

¹CI = Confidence Interval

Table S23. An OLS regression predicting the absolute difference between each participant's occupational age estimate and the average ideal hiring age of this occupation according to participants in the control condition. This model controls for the image uploaded by each participant for each occupation (Male or Female), as well as whether the occupation was coded as male by the majority raters in the control condition. Fixed effects by occupation and participant are included.

Table S23 provides clear support for Main Hypothesis 3. Participants in the image condition provided estimates of the age of occupations that were closer to the average perceived ideal age of an occupation among those in the control condition *when they uploaded a male image of an occupation* (β [Male Uploaded Image] = 0.72, CI = [0.35, 1.10], $p = .0001$); this effect holds controlling for whether the category was coded as male by those in the control condition, which did not on its own predict the gap between perceived age and ideal age (β [Category Coded as Male] = -0.28, CI = [-1.8, 1.20], $p = .71$). This suggests that when participants' uploaded an image of a male for a given occupation, their age estimates were closer to the stereotypical ideal hiring age for this occupation.

Hypothesis	Hypothesis Wording	Result	Reference
Main 1	The prediction is that when participants upload images of occupations they believe depict women, they will report the average age of this occupation to be lower than when participants upload images of men.	Supported	Figure 2A Figure 2B
Main 2	The prediction is that when participants upload images of occupations they believe depict women, they will report the average age of this occupation to be lower than the average age that people associated with this occupation in the control condition (without exposure to images depicting the occupation).	Supported	Figure 2A Figure 2B
Main 3	When participants upload an image of a male for a given occupation, their estimate for the average age of this occupation is going to be significantly closer to the perceived ideal age of this occupation (reported by participants in the control condition) than when participants upload an image of a woman for the same occupation. This includes when controlling for whether participants in the control condition report viewing an occupation as male or female typed.	Supported	Table S23
Main 4	The average age that participants report for an occupation in the control condition will correlate positively with the rate at which participants upload male images of this occupation (both in this current experiment and in past observational/experimental data collected with a similar paradigm, see preregistration: "Effect of Communication Modality on Strength of Gender Stereotypes (Replication + Extension 2)."	Supported	Figure 2C; S17
Main 5	The average age that participants report for an occupation in the control condition will correlate positively with the gender that	Supported	Figure 2C; S17

	participants most associate in their self-report beliefs (recorded via a slider) with each occupation as collected in past observational/experimental data collected with a similar paradigm, see preregistration: “Effect of Communication Modality on Strength of Gender Stereotypes (Replication + Extension 2).”		
Supp. 1	Participants’ ratings of hireability will positively correlate with their ratings of the average age of each occupation.	Rejected (ratings of hireability negatively correlate with age)	Figure S18
Supp. 2	Consistent with the theory of gendered ageism, we will also examine whether older women are more penalized than older men in terms of their perceived hireability. In other words, we will test whether the extent to which age is positively correlated with hireability is significantly stronger for men than women; our priors are not strong on this, but we will also test whether the correlation between age and perceived hireability may be entirely absent for women (but present for men), or whether it may even be reversed for women.	Partial (We find significant support that older women are more penalized than older men in perceived hireability; but we weakly expected a positive correlation between age and hireability which was mistaken; see supp. 1)	Figure S18

Table S24. A summary of all hypotheses (main and exploratory) explicitly stated in the preregistration for our Google Image search experiment. This table provides the exact wording of the pre-registered hypotheses, as well as whether our findings support or reject this hypothesis, and lastly, the figure or table that presents the results of the analysis corresponding to each hypothesis.

For the sake of comprehensiveness, we conclude by providing a table laying out each of our main and exploratory (supplementary) hypotheses, while indicating which were supported and which were not, as well as the corresponding figure or table that provides the results of the preregistered analyses. As Table S24 indicates, all of our main hypotheses were supported.

We also preregistered two exploratory supplementary hypotheses. These concern the exploratory outcome variable of ‘perceived hireability’ as measured by participants’ use of a Likert-scale (1 to 5) to indicate the perceived hireability of the image of the person they uploaded (i.e., these hypotheses related only to participants in the Image condition, since control participants did not upload images of people in occupations). Supplementary hypothesis 1 expected a positive correlation between participants’ hireability rating and the average age estimate of people in this occupation. Instead, we found the opposite. For women, this trend is significantly negative, and for men, it is slightly negative (Fig. S18).

Supplementary hypothesis 2 predicted that the relationship between age and hireability would be significantly different between images of women and men. We did not have strong priors, as our hypothesis wording indicates. We intended to run exploratory analyses to see if an interaction effect between gender and age was significantly predictive of perceived hireability. We went further and anticipated that the effect would specifically indicate that older women would be associated with lower hireability than older men. Our analyses provide statistically significant support for this claim. Figure S18 indicates this hypothesis as partially supported, because we weakly expected that older ages would be positively associated with hireability, which as our analysis of supplementary hypothesis 1 indicates, was not supported. Rejecting supplementary hypothesis 1 is slightly puzzling since we observe a strong positive correlation between the extent to which an occupation is rated as male and its perceived ideal hiring age averaged across participants in the control condition, and this correlation is highly robust and replicable across a range of measures of gender association (see Fig. S17). We further find that older ages are associated with higher resume quality scores by GPT (Figure 3). That said, given the robustness of our main predictions and results, we are hesitant to over-interpret these puzzling trends indicated by the Likert scale, especially in light of the known

limitations of Likert measures, which can be prone to noisiness, scale impression, extreme responding, and desirability bias²¹⁻²³. We think that exploring the methodological and psychological drivers underlying these surprising patterns is an interesting direction for future work.

Note: the framing of our study was informed by and updated in response to the review process for this project. Our initial submission emphasized what we referred to as the “invisibility of women online,” to highlight the underrepresentation of older women in our observational image and video datasets. However, as the reviewers correctly pointed out, our findings could also be framed to highlight the absence of younger men relative to younger women in these datasets. Moreover, the reviewers also pointed out that our analyses, such as Figure S18 (which was included in our initial submission), indicate that both women and men may experience discrimination as a part of age-related gender bias (in this case, Figure S18 suggests that younger women are preferred in hiring over younger men according to this Likert measure). For this reason, we revised our framing to highlight age-related gender bias as a statistical pattern that distorts underlying and measurable ground truth realities (e.g., in comparison to known age distributions of men and women throughout the workforce). Despite these framing changes, no changes were required to the framing of our experimental design of the statistical analyses undertaken or reported. None of the figures or results reported changed as a function of these framing revisions. Both our Google search experiment and our audit of ChatGPT were initially framed as an exploration of how exposure to online content via popular algorithms amplifies the statistical notion of age-related gender bias as identified in our online image and text datasets. For this reason, our pre-registered hypotheses – which focused on testing whether exposure to Google images amplified differences in people’s perceptions of the ages of women and men across occupations – continued to remain validly reported as consistent with our preregistration, even after revising the introduction of our study. This is indicated by the title of our preregistration – “Age-based Gender Bias and Its Amplification via online Content” – and is further confirmed by the detailed description of our intended study and tested hypotheses in the preregistration (Table S24).

E. 1. ChatGPT Audit Analysis

E. 1.1 Coherence of ChatGPT’s Responses

The main features of interest are highly correlated within each resume, indicating that ChatGPT generated coherent resumes based on the prompt. Fig. S19A shows that applicants’ age is highly correlated with applicants’ date of graduation ($r = 0.83$, $p = 2.2 \times 10^{-16}$), and Fig. S19B shows that applicants’ age is also highly correlated with applicants’ total years of relevant experience ($r = 0.84$, $p = 2.2 \times 10^{-16}$); accordingly, applicants’ total years of relevant experience is highly correlated with the years since their graduation ($r = 0.9$, $p = 2.2 \times 10^{-16}$), (Pearson correlation, two-tailed).

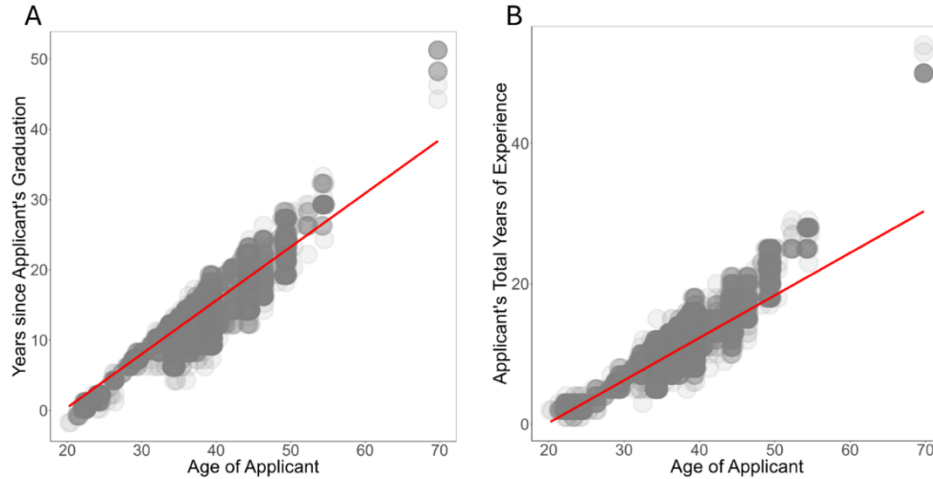


Figure S19. The raw correlation between the age of the applicant and the applicant’s (A) years since graduation and (B) total years of experience, as displayed on the resume generated by ChatGPT ($n = 36,720$ resumes total; 31,893 with reported years of experience). All data are from the treatment condition where ChatGPT was prompted with a specific name, either male or female. All data are shown across all names (i.e., across ethnic groups) and across all occupations. The trend line shows a linear trend produced by an OLS correlation with no control variables. Cases are removed from analyses if data from one of the relevant fields was missing in the resume generated by ChatGPT.

As further robustness, we tested whether these same correlations would hold when holding name and occupation constant and producing thousands of resumes to produce an overpowered sample of ChatGPT’s variation for this prompt. In this case, we set the temperature to the default level of 0.7. For this test, we selected at random a single name (“Chad Nichols”) and occupation (“programmer”) and generated 7,114 resumes for this specific prompt combination. Looking only at the resumes produced within this set, we see that applicants’ age is highly correlated with applicants’ date of graduation ($r = 0.97, p = 2.2 \times 10^{-16}$) and with applicants’ total years of relevant experience ($r = 0.78, p = 2.2 \times 10^{-16}$); moreover, within this fixed set of resumes, applicants’ total years of relevant experience is highly correlated with the years since their graduation ($r = 0.81, p = 2.2 \times 10^{-16}$). These results further demonstrate both the coherence and stability of ChatGPT’s responses to the prompt designed for our study.

E. 1.2 Comparing ChatGPT’s Age Assignments for Occupations to U.S. Census Data

As an additional test of the coherence of the resumes ChatGPT generated, we test whether the ages that ChatGPT generated for each occupation correlate with the median age of people in this occupation according to the U.S. Census from 2020 to 2023. For this comparison, we calculate the average age that ChatGPT generated for each occupation (averaging across both genders when gender is specified), and we compare this average age against the median age associated with this occupation during a given year of the U.S. census. We calculate these correlations separately for each condition in our resume generation experiment, and we control for condition, occupation, and census year using an OLS regression. The resulting model is presented below in Table S25.

Variables	Beta	95% CI ¹	p-value
Census Age (Median)	0.48	0.09, 0.87	0.014
Census Year			—
2020	—	—	—
2021	-0.08	-0.28, 0.12	0.43738
2022	-0.26	-0.52, 0.01	0.05600
2023	0.06	-0.17, 0.28	0.64306
Statistic	45.8		—
R ²	0.205		—
Adjusted R ²	0.201		—
No. Obs.	714		—
df	4		—
AIC	3,616		—
Residual df	709		—
Sigma	3.03		—

¹CI = Confidence Interval

Table S25. An OLS regression predicting the average age across resumes generated for an occupation by ChatGPT as a function of the median age of the same occupation according to the U.S. Census (2020 to 2023), while controlling for condition (in the resume generation task; see Extended Figure 2) and Census year, as well as clustering standard errors at the occupation category level.

Table S25 shows that the average age generated by ChatGPT for a given occupation across resumes is significantly correlated with the median age associated with this occupation in the U.S. Census in recent years ($\beta = 0.48$ years, CI = [0.09, 0.87], $p = 0.01$), while controlling for condition (in the resume generation task; see Extended Figure 2) and Census year, as well as clustering standard errors at the occupation category level. For reference, the raw correlation between ChatGPT’s average age associations and the median age in the census is $r = 0.45$ ($p = 2.2 \times 10^{-16}$, $t = 13.5$, Pearson correlation, two-tailed). These results indicate that ChatGPT’s age associations for occupations are consistent with underlying sociodemographic ground truth data on the median age of people in occupations according to the U.S. census, lending further credence to the coherence of its age representations.

E. 1. 3 Robustness to Model Temperature

In this section, we evaluate the robustness of the main results from the ChatGPT audit study to varying the temperature setting of ChatGPT when it generates its resume evaluation scores. The temperature feature is a hyperparameter that tunes the amount of variation and exploration that ChatGPT will accommodate in its outputs. Higher temperature means more variation (by analogy with entropy). Our main results assume the default temperature level in ChatGPT, which is 0.7. We examine the effects of both decreasing (to 0.3) and increasing (to 1.7) temperature. It is worth noting that increasing the model temperature leads ChatGPT to occasionally produce incoherent responses (such as providing scores above 100 when asked to provide a score between 1 and 100). This

occurred in less than 2% of cases in the raw data under the 1.7 temperature condition; these incoherent values are removed from the analyses, though all correlations are robust to and mostly unchanged by their inclusion. When the temperature is set to the default (0.7) or less, ChatGPT never produced any scores outside of the bounds requested in the prompt.

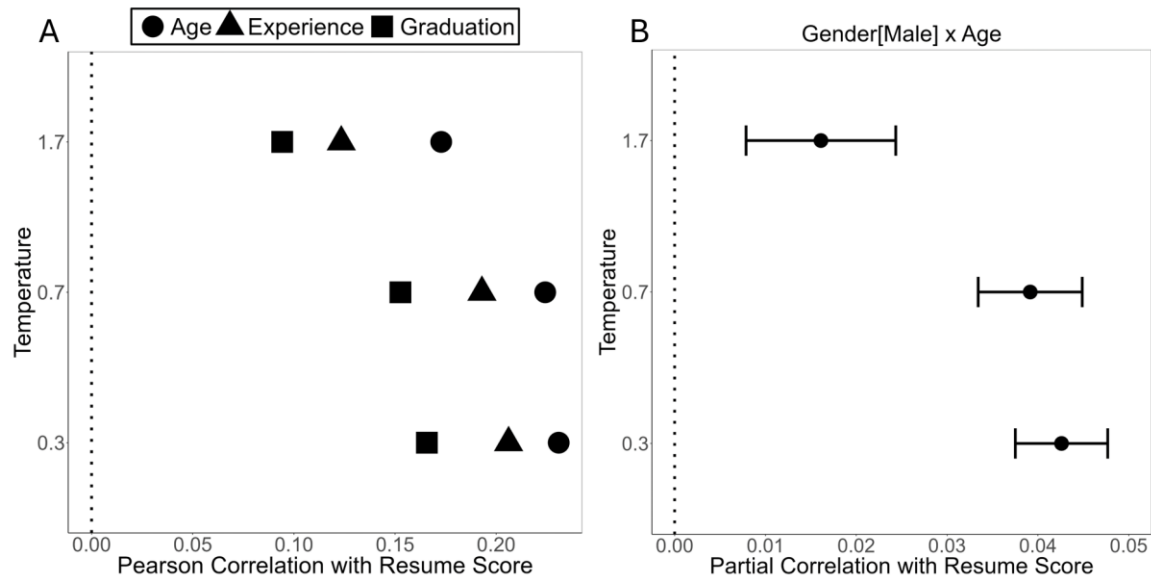


Figure S20. Robustness of correlation between outcome variables and ChatGPT's resume score while controlling for ChatGPT's model temperature. (A) The raw Pearson correlation of resume outcome variables (Age, Years of Relevant Experience, and Years since Graduation Date) and ChatGPT's resume scores, differentiated by model temperature. Resumes are excluded if ChatGPT did not include one of the relevant data fields (less than 8% of cases). (B) The partial correlation between ChatGPT's resume scores and an interaction term between resume Gender[Male] and Age, differentiated by model temperature. Correlations derive from an OLS model that controls for total years of applicant experience, the name of the applicant, and the target occupation (only data from the treatment condition is included in the analysis). Error bars display mean values \pm standard errors. Vertical dashed line indicates no correlation along the horizontal axis. Analyses consist of 37,848 resumes per temperature level.

Panel A of Figure S20 shows that, first off, that all outcome resume variables of interest – Applicant Age, Years of Experience, and Years since Graduation – are positively correlated with ChatGPT's evaluation score of resume quality. All correlations shown reflect highly significant Pearson correlations at the $p < 0.0001$ level. Importantly, panel A of Fig. S20 also shows that the significant positive correlation between all resume outcome variables and resume score is invariant to model temperature.

Panel B of Figure S20 similarly examines the robust of the interaction effect reported in the main text to varying model temperature. Specifically, each data point reflects the partial correlation between ChatGPT's resume scores and an interaction term between resume Gender[Male] and Age, differentiated by model temperature. Correlations derive from an OLS model that controls for total years of applicant experience, the name of the applicant, and the target occupation (only data from

the treatment condition is included in the analysis). The results indicate that, regardless of model temperature, there is a significant positive interaction between gender and age when predicting ChatGPT's resume evaluation scores; in other words, this means that, across all models, the positive effect of increased age on model scores is increased further if the applicant is male (consistent with the Figure 4C in the main text which shows that older men receive higher resume scores on average compared to older women, holding occupation constant). All correlations shown reflect highly significant correlations at the $p < 0.0001$ level.

Supplementary References

1. Rothe, R., Timofte, R. & Van Gool, L. Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks. *Int J Comput Vis* **126**, 144–157 (2018).
2. Ma, D. S., Correll, J. & Wittenbrink, B. The Chicago face database: A free stimulus set of faces and norming data. *Behav Res Methods* **47**, 1122–1135 (2015).
3. Alarifi, J., Fry, J., Dancey, D. & Yap, M. H. Understanding Face Age Estimation: humans and machine. in *2019 International Conference on Computer, Information and Telecommunication Systems (CITS)* 1–5 (2019). doi:10.1109/CITS.2019.8862107.
4. Kay, M., Matuszek, C. & Munson, S. A. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* 3819–3828 (Association for Computing Machinery, New York, NY, USA, 2015). doi:10.1145/2702123.2702520.
5. Metaxa, D., Gan, M. A., Goh, S., Hancock, J. & Landay, J. A. An Image of Society: Gender and Racial Representation and Impact in Image Search Results for Occupations. *Proc. ACM Hum.-Comput. Interact.* **5**, 26:1-26:23 (2021).
6. Gwet, K. L. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. (Advanced Analytics, LLC, Gaithersburg, Md, 2014).
7. Birhane, A., Prabhu, V. U. & Whaley, J. Auditing saliency cropping algorithms. in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* 1515–1523 (2022). doi:10.1109/WACV51458.2022.00158.
8. Nagpal, S., Singh, M., Singh, R. & Vatsa, M. Deep Learning for Face Recognition: Pride or Prejudiced? Preprint at <https://doi.org/10.48550/arXiv.1904.01219> (2019).
9. Buolamwini, J. A. Gender shades : intersectional phenotypic and demographic evaluation of face datasets and gender classifiers. (Massachusetts Institute of Technology, 2017).
10. Grother, P., Ngan, M. & Hanaoka, K. *Ongoing Face Recognition Vendor Test (FRVT)*. <https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing> (2019).
11. Padilla, R., Filho, C. F. F. C. & Costa, M. G. F. Evaluation of Haar Cascade Classifiers Designed for Face Detection. *International Journal of Computer and Information Engineering* **6**, 466–469 (2012).
12. Hassaballah, M., Murakami, K. & Ido, S. Face detection evaluation: a new approach based on the golden ratio Φ . *SIViP* **7**, 307–316 (2013).
13. Yang, B., Yan, J., Lei, Z. & Li, S. Z. Fine-grained evaluation on face detection in the wild. in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* vol. 1 1–7 (2015).
14. Guilbeault, D. *et al.* Online images amplify gender bias. *Nature* 1–7 (2024) doi:10.1038/s41586-024-07068-x.
15. Lilleberg, J., Zhu, Y. & Zhang, Y. Support vector machines and Word2vec for text classification with semantic features. in *2015 IEEE 14th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC)* 136–140 (2015). doi:10.1109/ICCI-CC.2015.7259377.

16. Pennington, J., Socher, R. & Manning, C. GloVe: Global Vectors for Word Representation. in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 1532–1543 (Association for Computational Linguistics, Doha, Qatar, 2014). doi:10.3115/v1/D14-1162.
17. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Preprint at <https://doi.org/10.48550/arXiv.1810.04805> (2019).
18. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching Word Vectors with Subword Information. Preprint at <https://doi.org/10.48550/arXiv.1607.04606> (2017).
19. Liu, Y. *et al.* RoBERTa: A Robustly Optimized BERT Pretraining Approach. Preprint at <https://doi.org/10.48550/arXiv.1907.11692> (2019).
20. OpenAI *et al.* GPT-4 Technical Report. Preprint at <https://doi.org/10.48550/arXiv.2303.08774> (2024).
21. Kreitchmann, R. S., Abad, F. J., Ponsoda, V., Nieto, M. D. & Morillo, D. Controlling for Response Biases in Self-Report Scales: Forced-Choice vs. Psychometric Modeling of Likert Items. *Front. Psychol.* **10**, (2019).
22. Jebb, A. T., Ng, V. & Tay, L. A Review of Key Likert Scale Development Advances: 1995–2019. *Front. Psychol.* **12**, (2021).
23. Kusmaryono, I., Wijayanti, D. & Maharani, H. R. Number of Response Options, Reliability, Validity, and Potential Bias in the Use of the Likert Scale Education and Social Science Research: A Literature Review. *International Journal of Educational Methodology* **8**, 625–637 (2022).