

Age and Gender Distortion in Online Media and Large Language Models

Corresponding Author: Professor Douglas Guilbeault

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)

Review of "The Invisibility of Older Women Online"

This paper explores the manifestation of age-based gender bias, arguing that women are consistently portrayed as younger than comparably aged men within internet content and among users of the internet. The authors offer evidence for this phenomenon via (a) publicly available images on the internet (e.g. Google Images), showing that women are consistently represented as younger than comparably aged men; (b) a nationally representative, pre-registered experiment, in which participants searched for google images of specific occupations and showed that participants rating these images similarly cast women as younger than men (and a greater willingness toward hiring younger women and older men); and (c) nine language models trained on billions of words on the internet showing that women and men are represented as younger and older, respectively.

REACTION

I felt this paper was far more impressive methodologically than theoretically. Methodologically I will commend the authors on this undertaking, comprising multiple large-scale datasets (albeit data that were used in prior published work, which some might argue reflects a problematic "double-dip"; see below). Nevertheless, for a journal as high-impact as Nature, I personally feel that an above-threshold contribution should reflect both method and theory. Unfortunately, in my view, and as per many of my critiques below, this paper falls way short in the latter domain. I would strongly urge the authors think carefully about what their work shows and which sub-literature(s) it truly adds to, because right now it's not clear.

MAJOR CONCERNS

-My biggest concern with this paper is ascertaining what precisely it is arguing, and why it matters. In other words, vis-à-vis contributing to the literature, what do we learn that we don't already know? (Or, alternatively - since I am not one who believes every paper should have a shocking, novel finding necessarily - confirming something we already know but in a way that really drives home the point by, say, introducing novel, transformative ways of testing a sound hypothesis.) Unfortunately, this paper does neither of these things, and instead, has me confused about what it does contribute.

First, I think it's well known that, for instance, women tend to be younger and/or valued for their youth more so than men. At least anecdotally, it has been widely known that the pressures toward childbearing and then childrearing has rendered older women "invisible" from a social role standpoint. For example, people often lament how women in Hollywood are seen as peaking more quickly than do men for prominent roles.

But even beyond that, the way in which the authors set up their argument seems overly confusing. For instance, the conclusion that "women are consistently represented as younger than men" does not appear to match the overall frame (including the title) of the paper that "older women are invisible." This implies that the focus is going to be on older women - i.e. (roughly) senior citizens who are female. But I think the more relevant conclusion would be that women of all ages are prescribed (pressured?) to be/act/seem younger, and that the authors show evidence of that phenomenon in multiple domains. This is a different argument than "older women are socially invisible." I don't see any of that in this paper, given the emphasis on "women are seen as younger" throughout the paper.

But even if the paper were somehow showing that “older women are more invisible than older men,” this is something that existing literature has already demonstrated in some form. Notably, I would urge the authors to check out the work of Martin et al. (2019, PSPB), who argue that older women reap an unexpected benefit of sorts by being “intersectionally invisible” – not prototypical of “older” or “woman,” so they are freer to act in assertive ways. This work is not only overlooked by the current authors, and not only helps refute the claim that “strikingly little is known about this dual bias,” but also presents a potential benefit for older women’s presumed invisibility, relative to older men. This further causes confusion in my mind about whether this paper is accurate in the first place in presuming bias. In other words, even if it is the case that women are consistently seen as younger than comparably aged men, is this always necessarily a bad thing? Do the authors truly show that this is a “bias” that necessarily disadvantages women (other than the hiring DV tossed into the experiment)? At the very least, the current findings do not do much in my mind to contribute to this subliterate on age-gender perception and related inequalities.

Let’s say that the authors were to reframe the paper to argue that “women are consistently represented as younger than men” as the main news (which is what I would strongly recommend). In this case, too, I would not be convinced this is news worth writing home about. First of all, this seems like a main effect, not an age-x-gender effect; i.e. women of all ages are consistently seen as younger than comparably aged men, regardless of age, so age does not seem to be a significant moderator here. Moreover, what does this “youth bias” mean? Why does it exist? Does it matter? Does it really mean there is a relative “invisibility” of aging among women? Is it due to physical differences, social pressures, role theory, or something else? And as per my prior comment, is being seen as younger, as the authors assume, truly a disadvantage? Way more theoretical clarity, and heft, needs to be added to this paper to make it worth publishing in a journal like Nature – or IMO any other high-impact journal, for that matter.

And in a similar vein, if this difference in perceived age really does come down to being an “age” effect, then how would we know this is not due to something related to age but not age itself, such as the well-known (or at least widely believed) phenomenon that women face greater expectations for looking attractive?

SMALLER POINTS

-I found the Abstract quite unclear, rather than clarifying at the outset what the paper is expected to deliver. What is meant by the statement that women are “consistently represented as younger than men” (represented how? measured how?); “searching for google images...amplifies bias in people’s beliefs...” (amplifies how? Measured how?); where do the “billions of words from the internet” come from? I do applaud the authors’ ambition and attempt at using a wide variety of big data sets, but IMO the Abstract needs to clearly set the stage for what is to come.

-In line with the confusing framing, I don’t understand why the authors go from arguing that “women are pressured to be young” (line 23-24) to talking about a gender pay gap that is more present among older women (26-27). There are several explanations for why this may be the case that have nothing to do with expectations of youthfulness, particularly if this youthful prescription (and a gender pay gap) applies to women on the younger wide, as well. Similar critique applies to the confusing transition from talking about the infantilization of women in academia (38) to then talking about the invisibility of older women online (40). These particular non sequiturs dovetail with my broader critique that the authors seem a little vague on what their paper is actually contributing, and why it matters. Much more work needs to be done to guide the reader to care about this work and see what it contributes.

-Although I recognize that the authors have done considerable legwork building their set of stimuli, the fact that the image dataset was not originally obtained with age in mind seems like it might lack a systematic age balance. Is there evidence that, for instance, all age/gender groups are equally representative, normed for perceived qualities like warmth, competence, and attractiveness, and the like? Perhaps the authors provided this info somewhere, but I don’t see it in the main manuscript nor the supplemental materials. Generally when dealing with facial stimuli, accounting for these potential confounds is critical.

-Discussion [544 etc.]: Again, not clear that the current studies “expose the invisibility of older women online.” This seems like a misrepresentation of the what the findings show, which is that women (of all ages) are consistently rated as younger than comparably aged men. The Discussion in general is full of the same issues that plague the paper’s framing/contribution as a whole, including the implication that this “bias” is automatically detrimental in all cases.

(Remarks on code availability)

Referee #2

(Remarks to the Author)

This is a really excellent manuscript. This research is very robust and the combination of methods is brilliant. This work not only evidences the case for the intersection of ageism and sexism, its a model of how to clearly and transparently draw together new AI methods and more traditional psychological experimentation to examine how the accumulation of biases (whether by human or artificial raters) reproduces structural cultural biases. I have no recommendations for this paper.

Peter Hegarty

(Remarks on code availability)

Referee #3

(Remarks to the Author)

This manuscript provides extensive and convincing empirical evidence on the underrepresentation of older women' in the online sphere. The authors perform an impressive exercise across a range of textual and image datasets, showing that women are consistently represented as younger online compared to men. In addition, they show how generative LLMs replicate these age- and gendered patterns once asked to generate CVs. Finally, authors show, in a set of neatly designed experiments, that the bias in gender and age representations in Google images search results amplifies biases in participants' beliefs about ages and genders of people in different occupations.

This is without a doubt a very admirable undertaking, and I highly commend how thorough and detailed the authors' analyses are. The manuscript presents compelling evidence of biases in a world that increasingly relies on image-based content and generative LLMs. As such, it is valuable to the general scientific community and the public alike. I also commend the authors for sharing their underlying data and well-documented analysis code.

In addition, I am glad to see a thorough, large-scale test of concepts (i.e., gendered ageism) we consider well-established in social sciences, but only have scattered evidence of. I found the empirical evidence thorough and rather convincing.

That being said, I am listing some notes and questions:

A. I think this manuscript delivers two important messages: (1) There are inequalities in reality that are reflected in the online sphere (shown by the image and LLM analyses); and (2) our perception of social reality is further distorted by the inequalities reflected in the online sphere (shown by the experiment).

Potentially, there is also an intermediate message: (1b) the inequalities in the online sphere are more severe than those in the society, which is once again amplified by (2). Authors suggest that "such age-based gender bias might systematically distort underlying sociodemographic realities". Yet, we are not told to what extent the biases authors identify mirror or deviate from the social realities. If most of the results mirror the social reality of gender composition of different occupations (e.g., the ones from the WordNet categories used in Google image searches and GPT-2 evaluations), the interpretation of most of these findings is somewhat different than if biases present in the society are *also* amplified in the online sphere before biasing individual decisions. I would strongly encourage authors to be explicit about this.

Authors do use actual census-based shares of men and women in different occupations in their robustness checks, but only to check for correlations with their alternative measures. I wondered why they did not take a step further to show the extent to which the online-content-based estimates reflect the reality of at least the US census (even though online content is not only curated based on content from the US and other English-speaking countries).

I found that the current organization of the paper does not emphasize these two arguments; I think making them more explicit could benefit the paper. Especially the placing of the experimental results – while obviously connected to the analyses of bias in images – created a discontinuity between these two arguments. Rearranging the results could improve the flow of the paper and make the relationship between the two messages of the paper more salient.

B. Whereas it is clear that the authors wanted to provide a comprehensive overview across various occupations/categories, I found the absence of any discussion of gender/age biases in relation to occupational status/prestige somewhat surprising. We know from the literature that women face a "glass ceiling" when it comes to entering prominent high-status positions (1). Such positions are occupied by older people; as women do not enter these positions, these older people are consequentially men. We also know that underrepresentation of certain groups in positions of power (2) or domains deemed to be of public and journalistic interest (3, 4) then leads to the these groups' underrepresentation in the public sphere. I would be surprised if text and images online (which both largely build on past and current media content) would show different regularities.

Thus, the online content is not necessarily encoding (only) the age-based bias against women, but the broader structural inequalities that cause women's absence from high-status positions – and, consequentially, online image and text content. Authors treat all occupations/categories in their datasets as equal, but I see this as a missed opportunity to reveal something more profound about the structural mechanisms behind the inequalities they observe.

Thus, I think that the manuscript captures the gender-age nexus (e.g., secretaries are younger women, chairmen of the board are older men), but conceals how much of the bias is actually driven by the gender-status inequality. E.g., figure 3 highlights "cook" as associated with young women; but I would assume that "chef", a higher-status "equivalent", would be associated with more senior men. Then, the effects of gendered ageism (which seems to be the emphasis of the manuscript) and structural gender discrimination (which results in gender inequalities at higher-status levels, but is not necessarily directly a result of ageism per se) are confounded. This might not be of utmost importance in making the general point this paper is trying to make. Yet, providing a more solid understanding of potential mechanisms underlying the empirical findings would, in my opinion, make the conclusions more comprehensive and convincing.

C. Related to Point B, having gone through the Google image search dataset, I had some questions/notes for the authors. I think the notes below also hold for the language model evaluations using WordNet categories.

Even when it comes to what authors denote as “ungendered” searches, some categories are clearly gendered. The most numerous categories related to women indicate youth by their nature (e.g., debutante, bachelorette, foster daughter, cheerleader, chorus girl, girl scout, check girl, working girl, brat, female child, princess, showgirl, valley girl). On the other hand, many of the categories for which many google images were evaluated for men clearly relate to very senior positions (e.g., cabinet minister, united states president, Mormon, government minister, civil rights leader, founding father, military volunteer, billionaire, civil leader, basketball coach, chief of state). Of course, the sole fact that so many women-related categories in WordNet are also youth-coded tells us a lot about the societal biases related to gender and age. However, one could not realistically expect to find older women online by searching for “chorus girls” or “cheerleaders”. Are the authors confident that the age/gender gap they find does not largely stem from the selective nature of the words that were searched for in the Google dataset? This goes back to my point on the age/status nexus. Do results change if one drops words that are both female- and youth-coded? If they do – or do not – can this tell us something about the origin/nature of these gender/age biases?

On a more pedantic note, checking the age estimates for very gendered words (daughter, trophy wife, girl scout, half-sister), I saw that the results for these categories depicting men were evaluated as older than those depicting women. How meaningful are these averages? How much of an impact do they have on the final results? I suspect not much would change if such categories were excluded, but I found such observations rather difficult to grasp. Some other categories made rather little sense overall (e.g., ape-man or Capricorn).

D. I found the abundance and current presentation of different analyses and results somewhat overwhelming. By the time I reached “Measuring Bias in AI applications”, I had trouble recalling the first analysis, and had to step back to understand how this relates to all the other results. Perhaps the authors can find a way to ease the mental load of going through so many diverse, yet related results.

I would also encourage the authors to make it clearer how the resume generation analysis complements and surpasses the results from the “Large-Scale Language Data” analysis. It is hardly surprising that, when asked to generate text based on their representations, LLMs reproduce the biases that are already contained in these representations. I would like a stronger emphasis on why we need both these analyses, presented separately, to gain a more complete image of the age-gender bias captured by the LLMs.

E. Figure S9 in the supplement highlights a finding I find interesting, but which remains neglected as it is not fully aligned with the broader narrative: in the experimental results, young women are considered more hireable for same positions compared to young men. Could it be that younger women are deemed more competent for some positions than young men? The situation then reverses for older people. I am once again wondering to what extent the job status – and the perception that there are no women in such high-status jobs – moderates this shift. That is, even though negative stereotypes surrounding women’s ageing prevail very broadly, and particularly in some industries (5), I wonder how much of the effect is driven by women’s absence from high-status categories.

References

1. D. A. Cotter, J. M. Hermsen, S. Ovadia, R. Vanneman, The Glass Ceiling Effect. *Soc. Forces* 80, 655–681 (2001).
2. E. Shor, A. Van De Rijt, A. Miltsov, V. Kulkarni, S. Skiena, A Paper Ceiling: Explaining the Persistent Underrepresentation of Women in Printed News. *Am. Sociol. Rev.* 80, 960–984 (2015).
3. P. England, S. Li, Desegregation Stalled: The Changing Gender Composition of College Majors, 1971-2002. *Gend. Soc.* 20, 657–677 (2006).
4. E. Shor, A. Van De Rijt, B. Fotouhi, A Large-Scale Test of Gender Bias in the Media. *Sociol. Sci.* 6, 526–550 (2019).
5. A. E. Lincoln, M. P. Allen, Double Jeopardy in Hollywood: Age and Gender in the Careers of Film Actors, 1926?1999. *Sociol. Forum* 19, 611–631 (2004).

(Remarks on code availability)

I have only ran a portion of the code; but I found it easy to read and follow. I also succeeded in producing several figures in the manuscript.

The specific descriptions of individual variables and the details about the information contained in each dataset were less clear.

Referee #4

(Remarks to the Author)

Summary:

Using large-scale image and text data the present work showed a persistent bias to represent women as younger than men online. Furthermore in an experiment, participants prompted to use a search engine for an occupation (vs. to search for control concepts) showed an enhanced female-young bias.

I found the work to be original and significant. I also found it to be generally methodologically sensible. I reviewed the main text, SI, and preregistration. Below I offer questions, comments, and suggestions.

Comments:

The authors anticipated a key concern I had, that is, that their results reflect bias in the annotators/coders and not bias in the images. I was reasonably convinced by their efforts to rule out this alternative account. One comment, for A. 1. 4. it seems to me that the key analysis is whether the relationship (slope) between actual age and perceived age varies by image gender (i.e., the actual age x image gender interaction term in a regression predicting perceived age). I could not find this clearly reported in the SI. What am I missing?

Why is the manuscript framed around the invisibility of older women rather than the invisibility of younger men? See especially Figure 4C. It appears resumes for younger women were scored higher than for younger men, based on the fitted line at least. I believe that there might be good reasons to focus on the invisibility of older women (e.g., Figure 2B, the experiment lowered women's age estimate but did not raise men's).

I would have liked to see more focus on effect sizes throughout. For instance, the age-gender association appears to be twice as large in the non-gendered searches than the gendered searches (p. 4). What do the authors make of this? How should readers think about the size of these effects? Any benchmarks?

I believe Reddit contributors have been documented to be extremely male-dominated. (Same for other text, Wiki, news) I understand that the results (e.g., p. 9) replicate across training algorithms and corpora. But how does the overrepresentation of men as contributors to the training set(s) influence the conclusions?

Given current preregistration standards, I suggest adding a table or similar to the SI to succinctly summarise adherence to and/or deviations from the preregistration. On my read, the preregistration appeared to be largely in line with what was reported. But I encourage greater clarity on this for both hypotheses and analyses to make it easier for readers to assess.

Minor points:

- I'm not sure what "mere cultural aesthetic" means in sentence 1.
- I found Figure 2.C unclear. Ultimately I was able to understand it, but I had to reread the relevant paragraph on p. 8 several times. For instance, 'gender associations' is unlabelled making it unclear what the poles are.
- Word embeddings capture mere co-occurrence as well as higher-order co-occurrence. That is, two words can be similar in vector space despite co-occurring with each other infrequently due to their frequent co-occurrence with the same other sets of words. I had trouble squaring this with the description on p. 9 of the gender dimension the authors constructed, which seemed to imply that it only captured mere co-occurrence.

(Remarks on code availability)

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)

I was Reviewer 1 on the previous version of this manuscript. My biggest concerns revolved around the theoretical framing of the paper, especially the argument around "older women invisibility" and the surprising lack of emphasis on what I thought was the more striking pattern in the data, which was that women of all ages seem to be perceived as act/look younger than comparably aged men.

I really applaud the authors' attention to my critiques, and also like their reframe on comparing perceptions to ground truth vis-à-vis age-gender perceptions. I agree that age is a more "objective" metric that serves as a useful comparison between stereotypes and reality. Thus, I think the theoretical contribution of the paper is far clearer than it was previously. This reframe, compared with the additional analyses presented, convince me that this paper is worthy of publication.

My last remaining request is that the authors do a more explicit job of highlighting why the authors selected age as their "ground truth" objective criterion. The Abstract in my opinion does not make this clear, nor does the Intro. I will leave it up to the authors to decide how to do it, but in general, I actually think the authors made a stronger case in their point-by-point response about how age serves as a useful objective metric for comparing the accuracy of stereotypes. I would like to see this same level of care in the main text, as well.

Once again, I commend the authors on a convincing reframe, analysis, and important paper.

(Remarks on code availability)

N/A

Referee #3

(Remarks to the Author)

I appreciate the authors' thorough and thoughtful engagement with my and other reviewers' comments and suggestions.

I find the revised framing which now explicitly tackles the comparison between the ground truth and online bias, as well as the two types of bias amplification, more coherent and impactful. I was glad to hear that the authors found my suggestion on this comparison useful.

I do find the granularity of the US Census data less than ideal; at the same time, I am convinced that the authors provided insights on the ground truth of average ages per industry to the best of their abilities.

The additional analyses on status add depth to the discussion; I find the insight that the age gap persists despite controlling for different facts of status informative - both theoretically and empirically. I was impressed by the breadth and comprehensiveness of status-related analyses conducted by the authors.

These revisions comprehensively address my concerns, and I am convinced by the arguments in the authors' response to reviewers.

(Remarks on code availability)

Referee #4

(Remarks to the Author)

I reviewed a previous version of this manuscript. I read the author's response to my comments as well as the new version of the manuscript. I found this revision to be responsive to my comments and those of the other reviewers. I think an already strong manuscript is now clearer with the interpretation of the data better articulated. I have no further comments.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Reply to Reviewer 1 (R1)

Referee #1 (Remarks to the Author):

Review of “The Invisibility of Older Women Online”

This paper explores the manifestation of age-based gender bias, arguing that women are consistently portrayed as younger than comparably aged men within internet content and among users of the internet. The authors offer evidence for this phenomenon via (a) publicly available images on the internet (e.g. Google Images), showing that women are consistently represented as younger than comparably aged men; (b) a nationally representative, preregistered experiment, in which participants searched for google images of specific occupations and showed that participants rating these images similarly cast women as younger than men (and a greater willingness toward hiring younger women and older men); and (c) nine language models trained on billions of words on the internet showing that women and men are represented as younger and older, respectively.

REACTION

I felt this paper was far more impressive methodologically than theoretically. Methodologically I will commend the authors on this undertaking, comprising multiple large-scale datasets (albeit data that were used in prior published work, which some might argue reflects a problematic “double-dip”; see below). Nevertheless, for a journal as high-impact as Nature, I personally feel that an above-threshold contribution should reflect both method and theory. Unfortunately, in my view, and as per many of my critiques below, this paper falls way short in the latter domain. I would strongly urge the authors think carefully about what their work shows and which sub-literature(s) it truly adds to, because right now it’s not clear.

MAJOR CONCERNS

-My biggest concern with this paper is ascertaining what precisely it is arguing, and why it matters. In other words, vis-à-vis contributing to the literature, what do we learn that we don’t already know? (Or, alternatively - since I am not one who believes every paper should have a shocking, novel finding necessarily - confirming something we already know but in a way that really drives home the point by, say, introducing novel, transformative ways of testing a sound hypothesis.) Unfortunately, this paper does neither of these things, and instead, has me confused about what it does contribute.

- We thank R1 for their deep engagement with our study and for pushing us to clarify our contributions. We have heavily revised our main framing and have added extensive, new data and analyses to address your concerns. We provide more detail on these revisions in reply to comments below.

In short, we no longer orient our study around the “Invisibility of Older Women,” since as you highlight, this captures only a part of our findings. Instead, we contextualize our findings with respect to a longstanding and ongoing debate concerning whether culture-wide stereotypes are accurate (by capturing observable correlations among the features

of social groups) or socially distorted (reflecting exaggerated or even illusory representations of the social world). The study of age-related gender bias is a particularly rich context for exploring these tensions given that age representations can be compared to ground truth data on the actual age of women and men throughout society. As our revised introduction explains (with extensive citations from the subliterate on stereotype accuracy, both in general and with respect to gender and age specifically), prior debate on this topic has been limited by the (i) lack of culture-wide multimodal data on the association between gender and age, and (ii) the lack of computational analyses comparing these associations to ground truth indicators. In our revised study, we have included extensive analyses that directly compare the representation of gender and age in online content to corresponding ground truth data on the age of women and men across industries (we copy these relevant additions below). Based on our extensive review of prior research on this topic, we can find no study that achieves these aims either partially or all together. We have compiled a table summarizing the relevant literature on this topic and highlighting the critical gap filled by our study. We believe these revisions effectively highlight how our study directly introduces, in R1's words, "*novel, transformative ways of testing a sound hypothesis*" – one of far-reaching consequences both methodologically and theoretically across the social sciences. Our newly added comparisons to census data are provided through Extended Data Figures 1-3, which we copy below for your direct review along with the associated prose.

Furthermore, our study not only allows us to test whether culturally widespread stereotypes (in this case, age-related gender bias) are accurate or distorted, but also related hypotheses about how stereotypical associations are amplified by the social status and prestige of occupations. We now test a foundational sociological hypothesis – namely that gender stereotypes are more salient for higher status, prestigious occupations – which has to date been observed only in small scale studies, and has yet to be tested in the context of age-related gender bias in large-scale internet data. We tested this in several ways. We conducted a nationally representative U.S. survey (N=1,002) examining people's judgments of the social status and prestige of the 867 occupations that could be matched between our observational datasets and the U.S. census. We also retrieved measures of occupational prestige from the U.S. Bureau of Labor Statistics, as well as the yearly earnings associated with these occupations. Across all measures, our results show that the age gap in gender representation (presenting men as older than women) is significantly higher for occupations associated with higher social status and wealth. This presents a rich sociological lens on our findings and provides yet another novel test of a sound hypothesis in the context of age-related gender associations online. These results are presented in Extended Data Figure 4, and we copy them in response to a different comment from R1.

Finally, in our revised framing, we argue that – by showing how age-related gender bias online is a systematic distortion that pervades the digital sphere – our study is uniquely poised to theorize and demonstrate the ways in which mainstream algorithms amplify

this distortion and thereby skew the social construction of how people represent the social world. This is what motivates our experimental analyses of how age-related distortions of gender are amplified by the Google Image search engine and by the use of ChatGPT in a popular workplace application to generate and evaluate resumes. We hope these newly added revisions and clarifications more effectively convey the novelty and importance of our findings in terms of both empirical and theoretical implications.

R1 may also be interested in the additional analyses of the U.S. census data provided by Extended Data Figure 1 and via supplementary materials further indicating that gender-age associations in online media are distorted relevant to the ground truth. These additional analyses demonstrate a lack of correlation over the past decade between the percentage of women in an occupation and the median age associated with this occupation in the U.S. census (Extended data figure 1; Fig. S1; Table S1). These revisions are also provided below for your direct review.

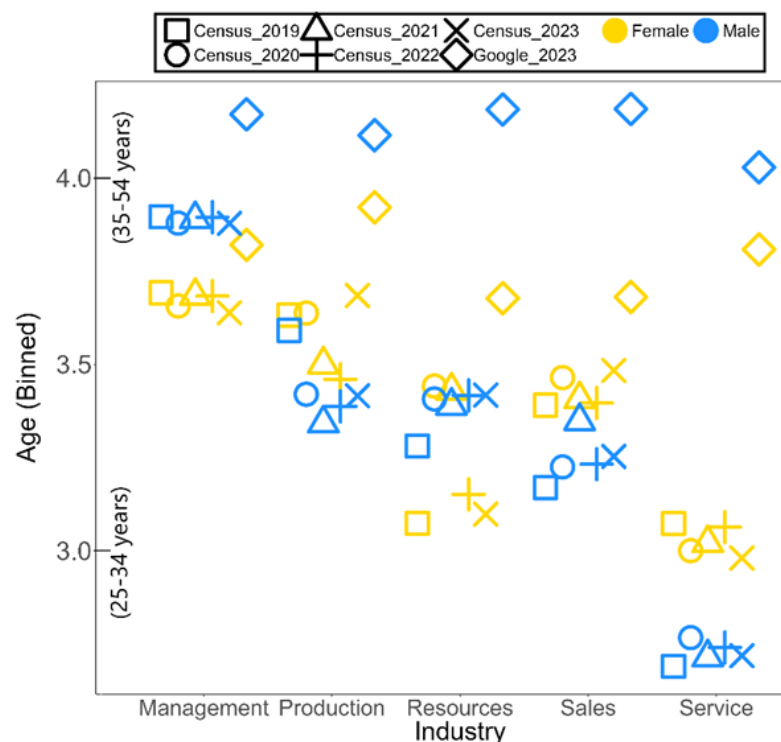
Lastly, for R1's direct review, we also provide Memo Table 1 summarizing recent papers on age-related bias and the key missing gaps covered by our study.

New analyses comparing against ground truth census data (pg. 7):

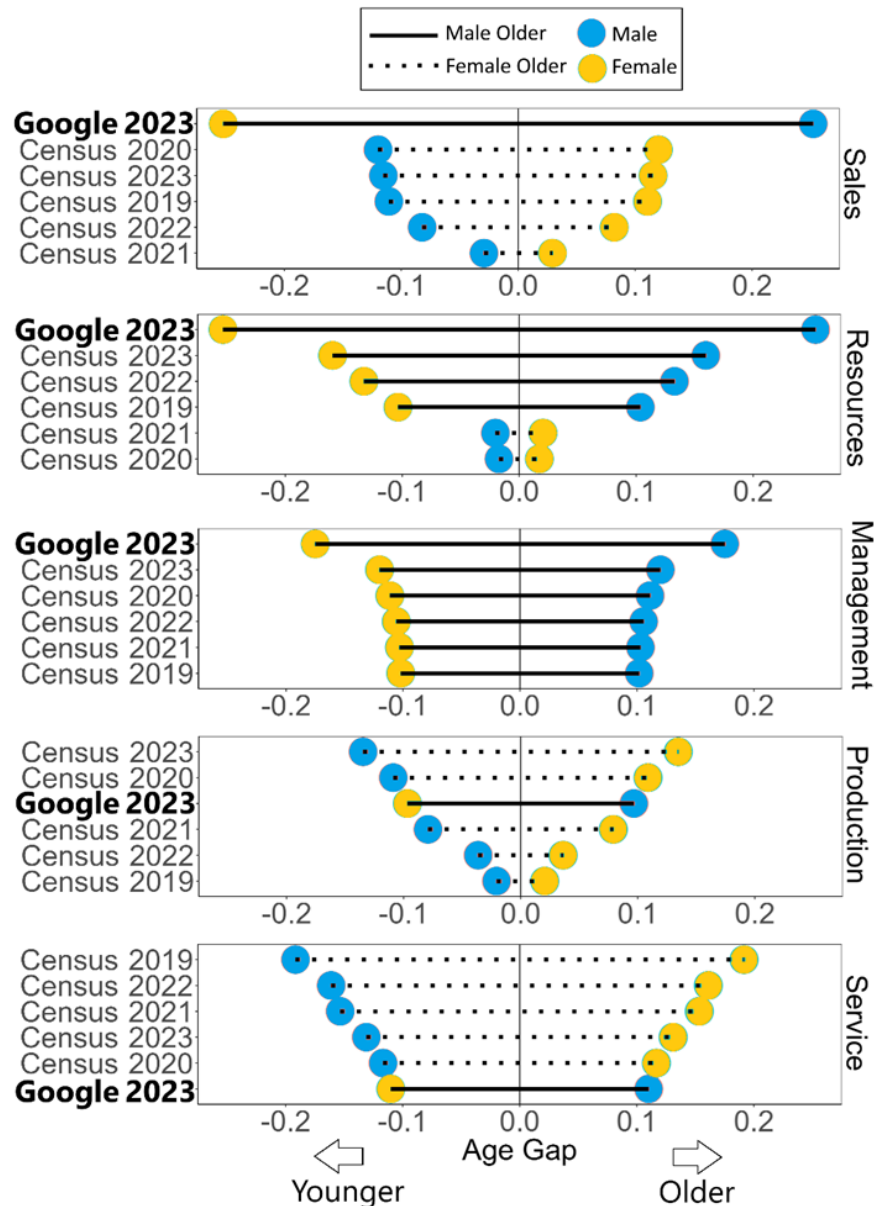
“Comparing to the Census

We now compare these findings to available ground truth data. We were able to match 867 social categories from our main Google image (Fig. 1A) dataset to occupational categories in the U.S. census. First, we observe that the average age rating for faces associated with occupations in Google is significantly correlated with the median age of people in these occupations according to the census (Tables S8 & S9; includes replications via Wikipedia). This indicates that the age estimates provided by the human coders are consistent with empirical age distributions across occupations. Yet, comparing to the census also shows how online images distort ground truth age-gender associations. The U.S. Bureau of Labor Statistics recently released a breakdown of the median age of each gender, from 2019 to 2023, across five industries: sales, services, natural resources and construction, production and transportation, and management. The census assigns each occupation to one of these industries, allowing those occupations matched in our Google image dataset to be assigned a census industry. We estimate the relationship between gender and age at the industry level by averaging the age associations in Google images across all occupations within a given industry (averaged within each occupation and then across occupations at the industry level). Conveniently, the census age groupings are highly similar to the age groupings the coders used when judging faces.

While gender-age associations in Google images and the census data are correlated at the industry level ($r = 0.13$, $CI = [0.11, 0.15]$, $p = 2.2 \times 10^{-16}$, Pearson Correlation, two-tailed; Extended Data Figure 2), Google images consistently display exaggerated and, in some cases, inverted trends that amplify the association between women and youth. Extended Data Figure 3 presents the absolute age gap between women and men by each industry, vertically ranked in terms of the magnitude of this gap, while also placing the older gender on the right side. In the industries of sales, resources, and management, Google images consistently present the highest age gap relative to all census years ($p < .001$ for all pairwise comparisons, Student t-test, two-tailed). Moreover, in each of these industries, Google images display men as older than women, while women are actually older than men for each of the census years examined in the sales industry, and for two of the years in the Resources industry. In the production and service industry, the magnitude of the age gap captured by Google images is not higher than all census years, yet the bias toward representing men as older is stable. In each census year examined, women are actually older than men in the production and service industry. It is only in Google images that men are older than women in these industries.”



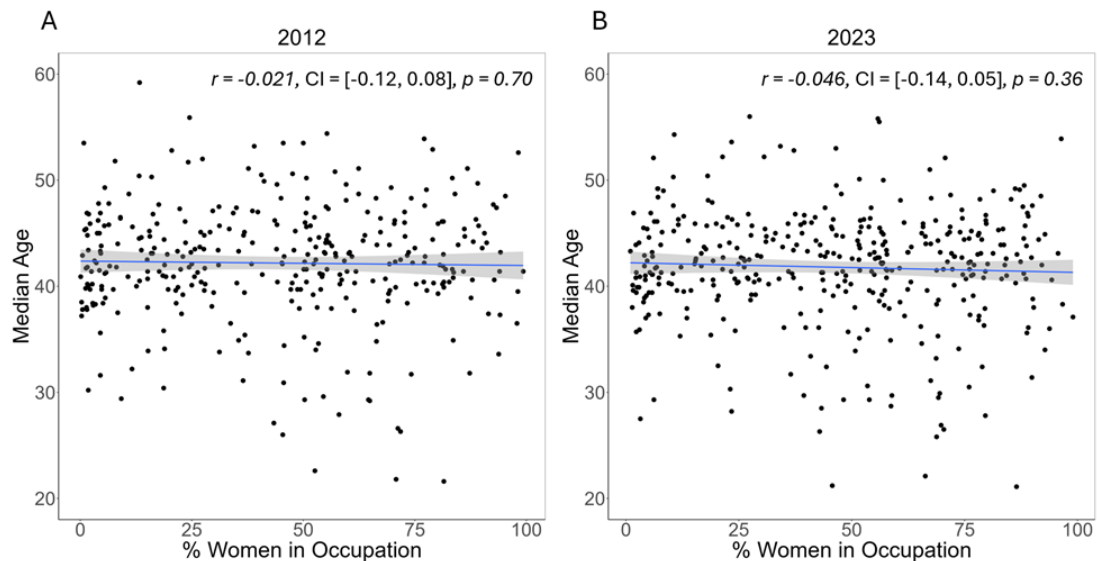
Extended Data Figure 2. Comparing the average age of women and men across industries in the U.S. Census (from 2019 to 2023) to the average perceived age of people in occupations from these same industries according to Google Images. The shape of the points indicates the data source, and the color of the points indicates the associated gender ($N = 867$ matched occupations).



Extended Data Figure 3. Comparing the magnitude and direction of the gender-age gap across industries in the U.S. Census (from 2019 to 2023) to the average perceived age of people in occupations from these same industries according to Google Images. Each panel shows the age gap between each gender for each industry separately. The midpoint of the age gap for each industry and data source is centered at 0 to help visually compare the magnitude of the age gap across datasets for each industry. Negative values along the horizontal axis indicate the gender that is associated with the lower age (relative to the midpoint), whereas positive values indicate the gender that is associated with the older age (relative to the midpoint). The color of the point indicates which gender falls on each side of the gender gap. Bold lines indicate cases where men are associated with a higher age than women for a given data source in a given industry; dotted lines indicate cases where women are older than men.

Related census analyses in the introduction (pg. 2):

“Yet, on the other hand, the statistical association between women and youth contradicts observable socioeconomic realities. Since the 1960s, women have consistently outlived men in the U.S. by as much as eight years, a gap that has been increasing^{38,39}. Census data on occupations presents similarly puzzling trends (Fig. S1). Over the last decade, there has been no correlation between the fraction of women in an occupation and its median age according to the U.S. census (Extended Data Figure 1; Table S1). There are also no clear differences in the age distribution of women and men throughout the workforce (Fig. S2). In fact, from 2009 to today, employed men in the U.S. have been more likely to be under the age of 40 than women, with no statistical difference in the fraction of employed women and men above 40 (Table S2)⁴⁰. Moreover, recent surveys fail to observe gendered ageism in certain organizational settings and even suggest that older women may be less impacted by stereotypes than older men^{41–43}. These inconsistent findings resonate with broader critiques against claims of enduring gender inequality, such as research showing declines in gender stereotypes over the last century in online text^{44,45}, as well as recent studies showing that hiring across industries increasingly favors women^{46–48}. This dissonant landscape raises the question of whether age-related gender bias is an organization or industry specific problem, or whether it is a culture-wide distortion that continues to reflect and contribute to systemic inequalities.”



Extended Data Figure 1. The correlation between the percentage of women in an occupation and the median age of people employed in this occupation according to the U.S. Bureau of Labor Statistics. Panel (A) shows the raw data (with each data point showing a single occupation) for 2012 (the correlation is non-significant; $r = -0.021$, $CI = [-0.12, 0.08]$, $p = 0.70$, Pearson Correlation, two-tailed, $n = 536$ occupations). Panel (B) shows the raw data (with each data point showing a single occupation) for 2023 (the correlation is non-significant; $r = -0.046$, $CI = [-0.14, 0.05]$, $p = 0.36$, Pearson

Correlation, two-tailed, $n = 594$ occupations). For all census years for which this data is provided in this format (from 2011 to 2023), there is not a single year with a statistically significant correlation between the fraction of women in an occupation and its associated median age (Table S1). Error bands show 95% confidence intervals.

Memo Table 1: A table summarizing the contributions and gaps of over twenty recent papers exploring age-related gender bias.

Paper	Summary	Sample	Year	Sample Size	Main Outcome Variable	In texts?	In images or videos?	In AI?	Impacts on Hiring decisions?	Compares to ground truth data (census)?
Post, C., & Byron, K. (2015). Women on boards and firm financial performance: A meta-analysis. <i>Academy of management Journal</i> , 58(5), 1546-1571.	Female board representation has a positive relationship with accounting returns (particularly in countries with stronger shareholder protections) but no significant relationship with market performance.	Companies from multiple countries across different time periods	2015	92,545 firms (spanning 140 studies)	Returns (ROA, ROE); Market success (Tobin's Q, stock returns)	No	No	No	No	No
Duncan, C., & Loretto, W. (2004). Never the right age? Gender and age-based discrimination in employment. <i>Gender, Work & Organization</i> , 11(1), 95–115.	Workplace age discrimination affects workers of all ages but is particularly pronounced for women.	Employees from a major UK financial services company	2004	1000 employees	NA	No	No	No	Yes	No
Rupp D. E., Vodanovich S. J., Credé M. (2005). The multidimensional nature of ageism: Construct validity and group differences. <i>The Journal of Social Psychology</i> , 145(3), 335–362.	Younger individuals and men had significantly higher ageism scores on the FSA than older individuals and women	Two samples of undergrad students from a public university in the southeastern United States.	2005	Two samples (N = 353; N = 201)	Ageism, via the Fraboni Scale of Ageism (FSA)s	No	No	No	Yes	No
Neumark, D., Burn, I., & Button, P. (2019). Is it harder for older workers to find jobs? New and improved evidence from a field experiment. <i>Journal of Political Economy</i> , 127(2), 922-970.	Experimental evidence shows that age discrimination in hiring against older job applicants, particularly older women, is stronger than previously documented.	Fictitious résumés sent in response to job ads in 12 cities across the United States, focused on administrative, sales, and security jobs.	2019	40,000 resumes sent from fictitious job seekers of different ages.	Callback rates for job apps.	No	No	No	Yes	No
Clarke, L. H., & Griffin, M. (2008). Visible and invisible ageing: Beauty work as a response to	Older women engage in extensive "beauty work" (cosmetic	Women aged 50-70 years living in Southern Ontario, Canada	2008	44 women	NA	No	No	No	Yes	No

ageism. <i>Ageing and Society</i> , 28(05), 653–674.	procedures, hair dyeing, etc.) as a response to gendered ageism, particularly in workplace contexts where they feel pressure to maintain a youthful appearance to remain professionally competitive.									
Chrisler, J. C., Barney, A., & Palatino, B. (2016). Ageism Can Be Hazardous to Women's Health: Ageism, Sexism, and Stereotypes of Older Women in the Healthcare System. <i>Journal of Social Issues</i> , 72(1), 86–104.	Ageist and sexist stereotypes in healthcare negatively impact older women's health outcomes.	Literature review and analysis of existing research on ageism and sexism in healthcare settings	2016	NA	NA	No	No	No	No	No
Macdonald, B., & Rich, C. (1983). <i>Look Me in the Eye: Old Women, Aging, and Ageism</i> .	Discusses the unique challenges faced by older women, emphasizing the compounded effects of ageism and sexism.	personal narrative with theoretical analysis rather than an empirical study	1983	NA	NA	No	No	No	Yes	No
Fernandez-Mateo, I., & Fernandez, R. M. (2016). Bending the pipeline? Executive search and gender inequality in hiring for top management jobs. <i>Management Science</i> , 62(12), 3636-3655.	Executive search firms do not discriminate against women in the hiring pipeline for top management positions, but rather the main source of gender inequality appears earlier in the process, specifically in the candidate pool composition.	Database from a large executive search firm containing information about candidates considered for executive positions in the UK between 2005 and 2009	2016	10,970 people, 2,250 job searches	Candidate progress through various stages of the executive search process: initial review, shortlisting, and client rejection or acceptance	No	No	No	Yes	No
Jyrkinen, M., & McKie, L. (2012). Gender, age and ageism:	Women managers in both Finland	Women managers aged 30-60 from	2012	25 subjects and 4 focus groups	NA	No	No	No	Yes	No

Experiences of women managers in Finland and Scotland. <i>Work, Employment and Society</i> , 26(1), 61–77.	and Scotland experience complex intersections of gender and age discrimination throughout their careers, with particular challenges around appearance expectations, career progression, and work-life balance.	Finland and Scotland								
Clarke, L. H., & Griffin, M. (2007). The body natural and the body unnatural: Beauty work and aging. <i>Journal of Aging Studies</i> , 21(3), 187-201.	Older women engage in extensive "beauty work" to maintain a socially acceptable appearance, viewing natural aging as "letting yourself go" while positioning beauty interventions as necessary maintenance work required to remain professionally and socially viable.	Women aged 50-70 years from Southern Ontario, Canada	2007	44 women; interviews	NA	No	No	No	Yes	No
Krekula, C., Nikander, P., & Wilińska, M. (2018). <i>Multiple Marginalizations Based on Age: Gendered Ageism and Beyond. In Contemporary Perspectives on Ageism.</i>	A chapter examining how age-based marginalization intersects with gender and other social categories, demonstrating that gendered ageism manifests as multiple, context-specific forms of disadvantage that affect both older and younger people.	Literature review	2018	NA	NA	No	No	No	No	No
Charlesworth, T. E., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of	Using word embeddings analysis of large language corpora, the study found remarkably consistent gender stereotypes	Large-scale language corpora including: Children's books and TV shows Adult	2021	65 million words across multiple corpora	Gender bias in word embeddings	Yes	No	Yes	Yes	No

more than 65 million words. <i>Psychological Science</i> , 32(2), 218-240.	across both children's and adults' language, with stereotypical associations being similarly strong and stable across different age groups and time periods.	books and movies Web-based text								
Edström, M. (2018). Visibility Patterns of Gendered Ageism in the Media Buzz: A Study of the Representation of Gender and Age Over Three Decades. <i>Feminist Media Studies</i> , 18(1), 77–93.	Over three decades, media representations have consistently marginalized older women, highlighting persistent gendered ageism, in Swedish media.	Swedish media content (news,, magazines, & television from 1982, 2002, and 2012)	2018	6000 people shown media In 1982, 2002, and 2012	NA	Yes	Yes	No	No	Yes
Itzin, C., & Phillipson, C. (1995). Gendered ageism: A double jeopardy for women in organisations. In C. Itzin & C. Phillipson (Eds.), <i>Gender, culture and organisational change. Putting theory into practice</i> (pp. 84–94). London: Routledge.	Women face a "double jeopardy" in organizations due to the intersection of gender and age discrimination, which creates compounded barriers to career advancement and equal treatment in the workplace.	Theoretical analysis with case examples from organizational settings	1995	NA	NA	No	No	No	No	No
Handy, J., & Davy, D. (2007). Gendered Ageism: Older Women's Experiences of Employment Agency Practices. <i>Asia Pacific Journal of Human Resources</i> , 45(1), 85–99.	Employment agencies in New Zealand systematically discriminate against older women through various practices including stereotyping, selective marketing to employers, and channeling them into lower-paid, temporary positions.	Older women (aged 45+) seeking employment through agencies in New Zealand, plus employment agency staff	2007	NA	NA	No	No	No	Yes	No
Frericks, P., Maier, R., & De Graaf, W. (2007). European pension reforms: Individualization, privatization and	European pension reforms focusing on individualization and	Analysis of pension reforms and policies in European countries, with particular	2007	Three countries' pension systems and reforms	Gender pension gaps (differences in pension	No	No	No	No	No

gender pension gaps. <i>Social Politics: International Studies in Gender, State & Society</i> , 14(2), 212–237. https://doi.org/10.1093/sp/jxm008	privatization tend to reinforce or even increase gender pension gaps by failing to account for gendered life courses and care responsibilities.	focus on Germany, the Netherlands, and Italy		(case study approach)	outcome between men and women)					
Ojala, H., Pietilä, I., & Nikander, P. (2016). Immune to ageism? Men's perceptions of age-based discrimination in everyday contexts. <i>Journal of Aging Studies</i> , 39, 44–53.	Older men generally perceive themselves as "immune" to ageism, often distancing themselves from age discrimination by emphasizing their own agency and by constructing ageism as something that primarily affects women or "other" men.	Finnish men aged 50-70 years	2016	67 thematic personal interviews with 23 middle and working class men aged 50-70	NA	No	No	No	Yes	No
Veresiu, E., & Parmentier, M.-A. (2021). Advanced Style Influencers: Confronting Gendered Ageism in Fashion and Beauty Markets. <i>Journal of Consumer Research</i> .	Older women fashion influencers engage in three key practices (defying chronological age, claiming fashion and beauty expertise, and becoming visibility entrepreneurs) to challenge gendered ageism in fashion and beauty markets while simultaneously reinforcing some age-based status hierarchies.	Advanced style influencers (women aged 50+) who are fashion and beauty content creators, along with their social media content, followers' comments, and related media coverage	2021	10	NA	Yes	Yes	No	No	No
Kwong See S, Heller R (2004) Judging older targets' discourse: how do age stereotypes influence evaluations? <i>Exp Aging Res</i> 30:63–73	Age stereotypes influence how people evaluate older adults' speech, with identical discourse being rated more negatively when listeners believed it came from an older speaker.	University students evaluating recorded speech samples	2004	148	Ratings of Discourse Quality	No	No	No	Yes	No

Calasanti, T., & King, N. (2015). Successful Aging, Ageism, and the Maintenance of Age and Gender Relations. In <i>Successful Aging as a Contemporary Obsession: Global Perspectives</i> .	Critiques the concept of 'successful aging,' highlighting how it reinforces age and gender norms, often marginalizing older women.	Literature review	2015	NA	NA	No	No	No	No	No
Radeke, M.K., Stahelski, A.J. Altering age and gender stereotypes by creating the Halo and Horns Effects with facial expressions. <i>Humanit Soc Sci Commun</i> 7, 14 (2020).	Facial expressions can significantly alter age and gender stereotypes through the halo and horns effects, where positive expressions (smiling) reduce negative stereotypes and negative expressions (frowning) increase them, particularly affecting perceptions of older adults and women.	Subjects recruited	2020	470; male (n=212) female (n=258)	Stereotype ratings based on facial expressions	No	Yes	No	Yes	No
De Sutter, Femke, and Sofie Van Bauwel. "Uncovering the Hidden Bias: A Study on Ageism in Hollywood's Portrayal of Ageing Femininities in Romantic Comedies (2000-2021)." <i>DiGeSt-Journal of Diversity and Gender Studies</i> 10.1 (2023): 18-34.	Hollywood romantic comedies portray aging women in stereotypical and marginalized roles, reflecting gendered ageism.	Hollywood romantic comedies (2000-21)	2023	15	NA	No	Yes	No	No	No
Barrett, A. E., Mimbs, H., Soulie, B., Bastow, S., Dominguez-Sandru, R., Michael, C., & Frost, M. (2025). Gray hair and pink slips: An analysis of Twitter responses to gendered ageism. <i>Journal of women & aging</i> , 37(1), 35-42.	Analysis of Twitter data shows widespread discussions and personal accounts of gendered ageism, particularly in professional settings.	Twitter posts & responses related to gendered ageism	2025	440	NA	Yes	No	No	Yes	No

First, I think it's well known that, for instance, women tend to be younger and/or valued for their youth more so than men. At least anecdotally, it has been widely known that the pressures toward childbearing and then childrearing has rendered older women "invisible" from a social

role standpoint. For example, people often lament how women in Hollywood are seen as peaking more quickly than do men for prominent roles.

- We agree that anecdotal evidence of this abounds (as our original and revised study emphasizes), especially in a select few contexts and industries like Hollywood. However, scientific studies of this phenomenon are often small-scale, sparse, and yield inconsistent findings across difficult to compare contexts. For instance, one of the only recent quantitative papers we could find that seeks to analyze empirical data on age-gender associations in popular media only examines the entertainment industry in Sweden and consists of less than 6,000 images (see Edström et al. 2018 in Memo Table 1). How this relates to the U.S. context, let alone the broader internet, both in images and in text, remains unclear. Otherwise, some studies have examined perceptions of gendered ageism in small-scale survey studies of particular organizational contexts and in terms of specific features (e.g., perceptions of agency), including the important study by Martin (2019), which we discuss further in reply to a different comment. On the whole, these studies have yielded inconsistent results, sometimes suggesting older women are less impacted by stereotypes. As our revised introduction emphasizes, this dissonant landscape of findings raises the question of whether age-related gender bias is a limited, organization or industry specific problem, or whether it is a culture-wide bias that permeates online media across modalities and contexts. To the extent that it is the latter, the critical question remains: is this bias accurate with respect to the ground truth distribution of men and women in society, and does the online representation of this bias through mainstream algorithms impact the beliefs and judgments of people and artificial intelligence?

No prior work on age-related gender bias is able to speak to these foundational questions of interest due to (i) the lack of large-scale, culture-wide, multimodal data on age-gender associations (and on the internet specifically), (ii) the lack of computational methodologies for grounding these age-related gender representations in corresponding ground truth data from the census, and (iii) the lack of experiments demonstrating the role of mainstream algorithms (e.g., search engines and large language models) in amplifying this bias *en masse*. Our study uniquely achieves all of these aims with unprecedented scale and clarity.

But even beyond that, the way in which the authors set up their argument seems overly confusing. For instance, the conclusion that “women are consistently represented as younger than men” does not appear to match the overall frame (including the title) of the paper that “older women are invisible.” This implies that the focus is going to be on older women - i.e. (roughly) senior citizens who are female. But I think the more relevant conclusion would be that women of all ages are prescribed (pressured?) to be/act/seem younger, and that the authors show evidence of that phenomenon in multiple domains. This is a different argument than “older women are socially invisible.” I don’t see any of that in this paper, given the emphasis on “women are seen as younger” throughout the paper.

- We thank R1 for emphasizing the lack of clarity in our original emphasis on the invisibility of older women. We hope that the extensive revisions to the framing and analyses in our paper (summarized at length in our replies above) have helped to alleviate this concern and more clearly communicate the novelty and importance of our contribution.

But even if the paper were somehow showing that “older women are more invisible than older men,” this is something that existing literature has already demonstrated in some form. Notably, I would urge the authors to check out the work of Martin et al. (2019, PSPB), who argue that older women reap an unexpected benefit of sorts by being “intersectionally invisible” – not prototypical of “older” or “woman,” so they are freer to act in assertive ways. This work is not only overlooked by the current authors, and not only helps refute the claim that “strikingly little is known about this dual bias,” but also presents a potential benefit for older women’s presumed invisibility, relative to older men. This further causes confusion in my mind about whether this paper is accurate in the first place in presuming bias. In other words, even if it is the case that women are consistency seen as younger than comparably aged men, is this always necessarily a bad thing? Do the authors truly show that this is a “bias” that necessarily disadvantages women (other than the hiring DV tossed into the experiment)? At the very least, the current findings do not do much in my mind to contribute to this subliteration on age-gender perception and related inequalities.

- We thank R1 for pointing us to this fascinating study and the broader discipline-specific debates it advances. This study by Martin et al (2019) finds in a controlled set of survey experiments (with a couple hundred participants in each study) that people are less likely to make agency prescriptions toward older women than older men, suggesting that older men may be more disadvantaged in terms of the agency implications of age-gender intersectionality. It offers an important corrective toward the narrative that age-related gender bias uniformly disadvantages women. This comment prompted us to communicate this point directly with Professor Martin, who is a colleague of one of the authors, to help us crystallize our understanding of how to best communicate the contribution of our findings to this literature. Professor Martin expressed enthusiastic agreement regarding the novelty and importance of our findings relative to this prior work and the studies that have been published since. This conversation – combined with your comments and those of the other reviewers – helped us identify a better way to contextualize our findings, which we summarize below.

A key point of novelty that emerged through these interactions and our careful revisitation of the literature is the emphasis not only on the large-scale multimodal nature of our findings (as emphasized in our prior submission), but also on the unique capacity to benchmark our findings in ground truth data on the actual ages of women and men throughout society. This provides an unprecedented opportunity to advance a longstanding debate regarding whether culture-wide stereotypes are generally accurate in reflecting real-world correlations in the attributes of social groups, or whether they are systematic distortions that exaggerate or even contradict measurable sociodemographic realities. As we discuss below, we have added extensive analyses that benchmark our

findings against measurable ground truth realities of age and gender distributions, and we find clear evidence that age-related gender bias is not only a culture-wide multimodal bias (permeating online images, videos, and massive, diverse corpora of online text), but also that it systematically distorts ground truth realities. Having identified these patterns at a large-scale in internet data, this naturally raises the question of whether mainstream algorithms (which learn from and mediate our consumption of online data) amplify this bias. This is where our Google Image search experiment and resume audit of ChatGPT become both empirical and theoretically rich in terms of demonstrating the increasingly powerful role that algorithms play in the social construction of reality regarding the widespread diffusion of distorted stereotypical associations.

Throughout our revision, we clarify that by ‘bias’, our study refers to a statistical bias (i.e., a skewed association between women and younger ages), and not a normative, evaluative bias (which is closer to the prescriptive judgments measured by Martin et. al. 2019). As our revised introduction explains, the concept of ‘stereotype’ evokes both statistical expectations and normative evaluations, and our study focuses on the former. While focusing on statistical associations limits our ability to explore specific prescriptive norms of interest in this subliterature (such as those relating to perceptions of warmth and agency), what we gain is the ability to carefully measure statistical patterns across modalities at scale and directly compare these to the ground truth statistical distribution of women and men across age groups in U.S. census data on occupations and industries. This exploration at scale is important both methodologically and theoretically. The question of whether common stereotypes are accurate requires a large-scale analysis not only of how these stereotypical associations are distributed in mass culture, but also how these associations compare to underlying ground truth data. We address this gap in the study of age-related gender bias, yielding important findings that build upon foundational work spanning psychology and sociology in the novel context of digital culture. We have copied the revisions below with especially relevant sections in bold:

Corresponding revisions to the introduction (starting on pg. 1):

“While few deny that stereotypes – generalizations about social groups^{15–17} – are harmful, a central question remains contested in social science: are common stereotypes generally accurate^{1–4} or socially distorted?^{5–8} **Some argue that commonplace stereotypes accurately capture observable aspects of social groups, otherwise they would not gain such widespread adoption^{1–4,18,19}. Yet, others argue that most stereotypes are exaggerated or illusory^{5–8}. Assessing stereotype accuracy is challenging because stereotypes involve not only statistical associations (e.g., expected correlations among the features of a social group) but also normative judgments (e.g., that one group is superior to another) for which there is no well-defined ground truth^{15–17}. Even for statistical associations, identifying ground truth is difficult. In some cases, this stems from disagreement on how to measure the ground truth, such as enduring debates over how to measure intelligence²⁰ (a heavily stereotyped characteristic^{21,22}).**

Yet, even when there is agreement on the relevant constructs, there is often a lack of large-scale, quantifiable cultural data for measuring pervasive stereotypical associations and comparing these to ground truth indicators. As a result, research on stereotypes often yields inconsistent findings, calling into question the pervasiveness and impact of these biases. In this study, we overcome these limitations in the analysis of age-related gender bias.

On the one hand, evidence abounds that older women face a dual bias at the intersection of gender and age. Policy reports^{11,12,23,24}, media coverage²⁵, and workplace interviews^{9,10,26} indicate that older women are discriminated against in hiring and promotion across industries (known as “gendered ageism”^{9,10,26,27}). This is related to a general statistical bias toward associating women with expectations of youth. From entertainment media to the workplace, women face persistent pressures to appear young, imposing a “beauty tax” with sizeable financial and time costs^{25,28,29}. This bias also manifests in everyday language. For instance, women in academia³⁰ and industry^{28,31–34} are much more likely than men to be referred to using infantilizing pronouns (e.g., ‘girls’), potentially undermining their perceived professional status. **These patterns suggest that age-related gender expectations may form a culture-wide statistical bias that influences how people view social roles and relationships across domains, echoing theories of gender as a central organizing frame in human perceptions of the social world and beyond**^{36,37}.

Yet, on the other hand, the statistical association between women and youth contradicts observable socioeconomic realities. Since the 1960s, women have consistently outlived men in the U.S. by as much as eight years, a gap that has been increasing^{38,39}. Census data on occupations presents similarly puzzling trends (Fig. S1). Over the last decade, there has been no correlation between the fraction of women in an occupation and its median age according to the U.S. census (Extended Data Figure 1; Table S1). There are also no clear differences in the age distribution of women and men throughout the workforce (Fig. S2). In fact, from 2009 to today, employed men in the U.S. have been more likely to be under the age of 40 than women, with no statistical difference in the fraction of employed women and men above 40 (Table S2)⁴⁰. Moreover, recent surveys fail to observe gendered ageism in certain organizational settings and even suggest that older women may be less impacted by stereotypes than older men^{41–43}. These inconsistent findings resonate with broader critiques against claims of enduring gender inequality, such as research showing declines in gender stereotypes over the last century in online text^{44,45}, as well as recent studies showing that hiring across industries increasingly favors women^{46–48}. **This dissonant landscape raises the question of whether age-related gender bias is an organization or industry specific problem, or whether it is a culture-wide distortion that continues to reflect and contribute to systemic inequalities.**

We argue that this uncertainty is fueled by the (i) lack of culture-wide multimodal data on the associations between gender and age, and (ii) the lack of computational methodologies for comparing these associations to ground truth indicators. To date, there have only been a handful of studies examining age-gender associations in small-scale surveys and interviews with

professional women^{9,10,14,41,42,49–51}, or in sparse, non-representative observational studies of particular industries, such as celebrities and athletes in entertainment media^{26,28,52–54}. Yet, failing to observe age-related gender bias in small samples of a few contexts does not indicate its lack of prominence on a culture-wide scale. Recent work in human cultural evolution finds that social biases in how people categorize the world frequently emerge only at scale^{55–57} and can manifest as exaggerated or even illusory beliefs^{58,59}. This suggests the alternative view that skewed associations between gender and age can emerge as a large-scale statistical bias that distorts underlying socioeconomic realities, despite inconsistencies across small-scale samples and contexts.”

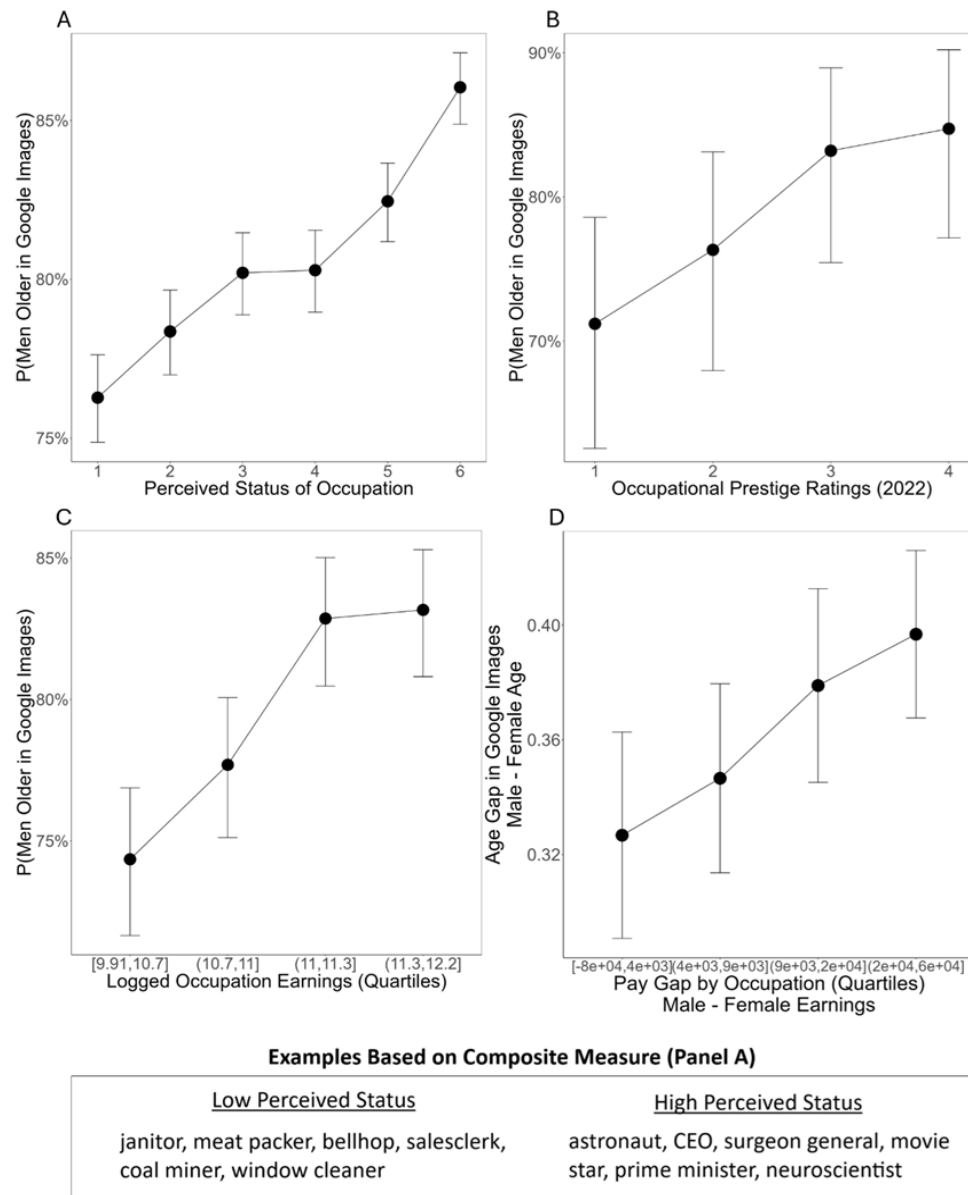
Our study further includes novel data and analyses that directly link age-related gender bias in our observational data to extensive indicators of socioeconomic inequalities among occupations. As we argue in our replies above, this provides novel support for foundational sociological theories that have yet to be tested at this scale in the context of age-related gender representations. We provide the relevant revisions and the corresponding Extended Data Figure 4 below.

“Relationship to Social Status

Given the observational nature of these analyses, it is challenging to evaluate specific mechanisms driving these large-scale age-gender associations. Nevertheless, numerous patterns in our data are relevant to considering possible explanatory factors. One such consideration pertains to the hypothesis that gender stereotypes are most salient in occupations associated with higher status and prestigious, since high-status occupations often receive the most collective attention and praise, thus playing a prominent role in reinforcing gender expectations and norms of desirability^{80–82}. In a follow-up study, we recruited a nationally representative sample of U.S.-based Prolific users ($N = 1,002$) to evaluate the status and prestige of 867 occupations matched between our main Google Image data (Fig. 1A) and the U.S. census from 2015 to 2022. We find that occupations rated as higher status are more likely to elicit Google images in which men are older than women (Extended Data Figure 4A; $r = 0.08$, $t = 11.28$, $p = 2.2 \times 10^{-16}$, Pearson Correlation, Two-Tailed, $N = 867$ Occupations). Examples of occupations in the lowest or highest 5% of perceived social status are provided at the bottom of Extended Data Figure 4.

We then go beyond people’s subjective perceptions of status (which can widely vary⁸³) by using ground truth indicators. First, we reproduce this correlation using the U.S. Bureau of Labor Statistics’ measure of occupational prestige⁸⁴ (Extended Data Figure 4B; $r = 0.11$, $t = 2.5$, $p = .01$, Pearson Correlation, Two-Tailed, $N = 532$ Occupations). Next, we test whether the median yearly earnings associated with occupations according to the U.S. census also predicts age-gender associations in Google images. Indeed, the probability of men appearing as older in Google images is significantly higher for occupations with higher earnings (Extended Data Figure 4C; $r = 0.11$, $t = 7.39$, $p = 1.07 \times 10^{-13}$, Pearson Correlation, Two-Tailed, $N = 4,444$ pairwise comparisons at the census-year level from 2015 to 2022; yearly earnings logged). We even find that the gender pay gap^{12,24} — the extent to which men earn more than women in the same occupation — is associated with the extent to which men appear older than women in Google images, i.e., with the digital age gap (Extended Data Figure 2D; $r = 0.04$, $t = 7.305$, $p = .002$, Pearson Correlation, Two-Tailed, $N = 4,444$

pairwise comparisons at the census-year level from 2015 to 2022; yearly earnings logged). These results are robust to a range of statistical controls (Figs. S11 & S12; Tables S10-13), and resonate with longstanding concerns regarding connections between economic disparities and disparities in how genders are perceived and evaluated in the workplace^{12,24}.



Extended Data Figure 4. The age gap for occupations in Google Images is predicted by the perceived status of occupations, as well as by the median yearly earnings of occupations and the gender pay gap by occupation according to U.S. Census data from 2015 to 2022. Google image data is from Guilbeault et al. (2024; see Fig. 1A) and is based on 866 social categories matched to occupations in the U.S. census. We recruited a nationally representative sample of U.S.-based Prolific users ($N = 1,002$)

to evaluate the status and prestige of 827 occupations matched between our main Google Image data (Fig. 1A) and the U.S. census. (A) The correlation between the perceived status of an occupation and the probability that men appear to be older than women in Google images of the occupation (status perceptions are averaged across a nationally representative sample of U.S. participants, $n = 1,002$ participants; an average of 27 participants rated each of occupations; data shown in six evenly spaced bins). See “Materials and Methods” for data collection and aggregation details. Examples of occupations in the lowest (highest) 5% of perceived social status according to this measure are provided at the bottom of the figure. (B) The correlation between the U.S. Bureau of Labor Statistics’ measures of occupational prestige (shown in quartiles) and the probability that men appear to be older than women in Google images of the occupation (532 occupations could be matched). (C) The logged median yearly earnings for an occupation (shown in quartiles) predict the probability that men appear to be older than women in Google images of the occupation. (D) The pay gap in median earnings for an occupation by gender (shown in quartiles) predicts the age gap in perceived age between men and women in Google images of the occupation. For (B) and (C), data are shown for the 753 occupations that could be associated with yearly earnings across Census years, 2015 to the present. Error bars show 95% confidence intervals.

Lastly, building on R1’s suggestion, we now discuss more explicitly how our findings highlight ways in which both women and men may be negatively impacted by age-related gender bias in the context of search algorithms and workplace applications of ChatGPT. We copy the corresponding revisions to our discussion section here:

“How might the mass distortion of age-related gender associations online negatively impact women and men? Our results highlight several key ways in which older women are likely to be disadvantaged by this bias. For example, when generating resumes, ChatGPT not only assumes that women are younger, but also that they have less overall experience; consequently, ChatGPT is biased toward giving lower scores to resumes from younger women compared to older women, while giving the highest scores to older men. Yet, ChatGPT also gives higher scores to resumes from young women compared to young men, suggesting that young men may also be disadvantaged by this dual bias, a finding also supported by some of our experimental measures of participants’ hiring preferences (Fig. S18). However, a selection bias favoring younger women and older men may further reinforce gender inequalities at the systemic level, whereby women are preferentially hired into roles with lower status and authority but denied mobility, while older men continue to enjoy their positions at the top. This resonates with our finding that online content is most likely to depict men as older than women for occupations with higher social status and wealth. Future research is needed to explore the ramifications of this dual bias on structural inequalities, especially in light of the sheer extent to which this bias permeates digital culture and mainstream algorithms.”

Let's say that the authors were to reframe the paper to argue that "women are consistently represented as younger than men" as the main news (which is what I would strongly recommend). In this case, too, I would not be convinced this is news worth writing home about. First of all, this seems like a main effect, not an age-x-gender effect; i.e. women of all ages are consistently seen as younger than comparably aged men, regardless of age, so age does not seem to be a significant moderator here. Moreover, what does this "youth bias" mean? Why does it exist? Does it matter? Does it really mean there is a relative "invisibility" of aging among women? Is it due to physical differences, social pressures, role theory, or something else? And as per my prior comment, is being seen as younger, as the authors assume, truly a disadvantage? Way more theoretical clarity, and heft, needs to be added to this paper to make it worth publishing in a journal like Nature – or IMO any other high-impact journal, for that matter.

- Thank you for this recommendation. As you can see in our response to your previous comment, you inspired us to move away from this main framing. Instead, inspired partly by what the other reviewers consistently identified as novel in our research, we decided to undertake extensive analyses and data collection to support a revised framing focusing on how large-scale associations between gender and age online systematically distort underlying sociodemographic realities (e.g. across occupations and industries) and how these distortions are amplified *en masse* by mainstream algorithms, i.e., the Google Image search engine and workplace applications of ChatGPT.

And in a similar vein, if this difference in perceived age really does come down to being an "age" effect, then how would we know this is not due to something related to age but not age itself, such as the well-known (or at least widely believed) phenomenon that women face greater expectations for looking attractive?

- Thank you for this insightful comment. It is not our goal to isolate whether the large-scale statistical biases we observe reduce to an age or a gender specific effect *per se*. On the one hand, our findings regarding the analysis of celebrities in visual content from IMDb, Google, Wikipedia, and Youtube is consistent with aesthetic pressures on women to look young. Yet, we also observe strikingly similar associations between women and youth in Google images of everyday social categories and occupations that often do not depict celebrities. Rather, these images come from across the web (with common sources being blogs and company websites; Fig. S10) and include real photographs from regular people, stock photos, avatars, and more. Even more to the point, the cultural pressure on women to appear young in visual content does not readily explain the robust, large-scale association between women and youth in our novel analyses of semantic associations among social categories and occupations across nine popular language models trained on billions of words. These semantic associations reflect large-scale and often implicit co-occurrence patterns between words and the linguistic contexts in which they appear (i.e., proximities among words in these embedding spaces are weighted not only by whether words appear together in shared contexts, but also whether they appear with similar words in similar contexts, even if not directly together). This empirical finding broadens prior theoretical work on age-related gender bias by indicating how it is shaped

not only by aesthetic pressures on visual appearances, but also by widespread patterns of verbal expression in daily language use on the internet. In our revised discussion, we emphasize how our findings highlight interesting new terrain for future research into the question of whether and how aesthetic norms of appearances and attractiveness in entertainment media spillover and shape how people conceptualize age and gender even in non-visual verbal expressions.

“ [...] a critical direction for future research is to investigate the causal mechanisms through which age-related gender bias seeps into and spreads via the images, videos, and text of distinct platforms, each with its own unique audiences and distribution channels. **For example, our results regarding objective differences in the ages of male and female celebrities visualized on IMDb, Wikipedia, and Google likely reflect industry-specific mechanisms relating to status dynamics, hiring biases, and the objectification of women in entertainment media. Yet, these industry-specific drivers do not readily account for how strongly women and youth are semantically associated in massive bodies of online text from diverse sources, let alone in ChatGPT’s text-based representations and rankings of job candidates. A fascinating question for future work is to explore the extent to which the aesthetic norms, fictionalized representations, and hiring biases of entertainment media spillover into the mass distortion of age-gender associations in other areas of social life.** A related question concerning supply-side factors concerns how age-related gender bias in popular algorithms may stem from inequalities in the gender of data contributors online; studies suggest that Reddit users⁹³ and Wikipedia editors⁹⁴ are disproportionately male, and textual data from these platforms are frequently mined for training AI models. Training AI on datasets with greater gender equality in data contributors may provide an effective mitigation strategy.”

SMALLER POINTS

-I found the Abstract quite unclear, rather than clarifying at the outset what the paper is expected to deliver. What is meant by the statement that women are “consistently represented as younger than men” (represented how? measured how?); “searching for google images...amplifies bias in people’s beliefs...” (amplifies how? Measured how?); where do the “billions of words from the internet” come from? I do applaud the authors’ ambition and attempt at using a wide variety of big data sets, but IMO the Abstract needs to clearly set the stage for what is to come.

- We appreciate this recommendation. We have revised our abstract to more clearly align with our revised framing and core findings, while also minding the strict stylistic length and formatting requirements of *Nature* abstracts.

-In line with the confusing framing, I don’t understand why the authors go from arguing that “women are pressured to be young” (line 23-24) to talking about a gender pay gap that is more present among older women (26-27). There are several explanations for why this may be the

case that have nothing to do with expectations of youthfulness, particularly if this youthful prescription (and a gender pay gap) applies to women on the younger wide, as well. Similar critique applies to the confusing transition from talking about the infantilization of women in academia (38) to then talking about the invisibility of older women online (40). These particular non sequiturs dovetail with my broader critique that the authors seem a little vague on what their paper is actually contributing, and why it matters. Much more work needs to be done to guide the reader to care about this work and see what it contributes.

- Thank you once again for emphasizing these important opportunities to improve the clarity of our work. We agree that our original framing surrounding “the invisibility of women online” was a limited and confusing way to summarize the full breadth of our results. We have since removed this framing from our introduction. We also agree that our prior attempt to discursively connect age-related gender bias in cultural data to economic research on the gender pay gap was confusing, so we have removed this from our introduction. We believe our revised framing and extended results clearly capture what our paper contributes and why it matters. Thank you once again for spurring many improvements to our writing through your insightful comments.

-Although I recognize that the authors have done considerable legwork building their set of stimuli, the fact that the image dataset was not originally obtained with age in mind seems like it might lack a systematic age balance. Is there evidence that, for instance, all age/gender groups are equally representative, normed for perceived qualities like warmth, competence, and attractiveness, and the like? Perhaps the authors provided this info somewhere, but I don't see it in the main manuscript nor the supplemental materials. Generally when dealing with facial stimuli, accounting for these potential confounds is critical.

- We sincerely appreciate this thoughtful comment and recognize the importance of clarifying some aspects of our study in response. Regarding our observational image analyses, for the Guilbeault et al. (2024) dataset, MTurkers' age judgments of the images were initially collected as part of the original dataset. The dataset was collected with the aim of providing a representative sample of images across all social categories available in Wordnet. Specifically, the data collection procedure was designed to be representative of how Google Images depict the ontology of social categories, such that any age-related or gender-related statistical biases that emerge at this scale are of interest in terms of capturing large-scale demographic dimensions skewing human representations across the full spectrum of social categories.

The additional datasets we examine were collected by other research teams and contain either ground-truth data on the true age of faces depicted, or machine learning classifications of gender and age, where this age information was provided by the original papers, though none of these papers (to our surprise) examine the correlations between gender and age in their own datasets. As emphasized in our paper, all of our results replicate strongly in these additional datasets, in terms of direction, magnitude and significance.

Moreover, we have taken multiple steps to ensure that differences in coders' perceptions along these dimensions are not likely to impact our results as a confound. Specifically, these steps include (i) the replication of our results across datasets and coding methods involving automated, machine-learning classifications rather than human judgments, and (ii) our robustness validation survey which shows that our human coders were equally capable of accurately judging the age of faces in a dataset with all age-gender groups equally represented and the true age of the faces is known. Notably, our coders did not indicate different accuracy rates in their judgments of women and men across age groups, which suggests that differences along the dimensions that R1 highlights are unlikely to pose a confound in our current results.

While we agree that having information about perceptions of warmth, competence, and attractiveness in online images could lead to interesting explorations, we believe this falls beyond the scope of our project, especially in light of the recommendation from the Editor and reviewers to streamline our already extensive analyses. Recent work by Sun et al. (2024) shows that measuring perceived attractiveness and other presentational, stylistic features in online images is challenging, largely due to the subjective nature of rating attractiveness and related qualities. We imagine perceptions of warmth and competence will face similar challenges, particularly in the absence of established machine learning classifications for comparison. Given these considerations, we have highlighted this as an important direction for future research in our revised discussion (copied below). We hope this discussion helps address R1's concerns. We greatly appreciate this insight that led us to be more specific in the descriptions of our data sources, methodological choices, and the robustness checks we conducted to ensure the validity of our findings.

Revision to the discussion (pg. 17):

“In this study, we provide large-scale evidence not only that age-related gender bias pervades online media – spanning images, videos, and texts across major platforms – but also that the bias toward representing women as younger systematically distorts the observable ground truth realities on the actual ages of women and men throughout society. Our findings raise alarm regarding the algorithmic amplification of age-related gender bias on a massive scale, especially considering that many mainstream machine learning algorithms are trained on these public datasets. Indeed, many of the image and text datasets that we examine in this study are used extensively as canonical training and benchmark datasets for developing AI applications. Enormous harm can be caused by latent social biases that lurk in popular machine-learning tools^{60,87,88}, and algorithmic bias typically arises from contaminated training data. In our study, we provide direct evidence that age-related gender bias is actively amplified through two of the most widely used algorithms today – the Google Image search engine and ChatGPT. While companies like Google and OpenAI invest heavily in reducing stereotypical content in their products^{89,90}, most work in this area focuses on

single dimensions of bias, such as gender-based or race-based biases. Our study highlights the critical need to account for multimodal and multidimensional forms of bias⁹¹, which are more challenging to detect but no less consequential in how people and algorithms represent the social world. **The intersectional statistical bias we identify between gender and age may ultimately interact with other biases – for example, relating to how women and men are depicted in terms of their warmth and competence – revealing a promising direction of future research that our methods are well-positioned to advance^{49,92}.**

References in revised passage:

49. Fiske, S. T. Stereotype Content: Warmth and Competence Endure. *Curr Dir Psychol Sci* **27**, 67–73 (2018).

92. Sun, L. *et al.* Smiling women pitching down: auditing representational and presentational gender biases in image-generative AI. *Journal of Computer-Mediated Communication* **29**, zmad045 (2024).

-Discussion [544 etc.]: Again, not clear that the current studies “expose the invisibility of older women online.” This seems like a misrepresentation of the what the findings show, which is that women (of all ages) are consistently rated as younger than comparably aged men. The Discussion in general is full of the same issues that plague the paper’s framing/contribution as a whole, including the implication that this “bias” is automatically detrimental in all cases.

- Once again, we appreciate this astute recommendation for how to improve our framing. We fully acknowledge that our initial framing may have overstated the idea of the “invisibility” of older women online. We have carefully revised our paper to ensure our findings are presented with greater precision. In alignment with your recommendation, we have removed references to this concept, since our results depict many more interesting aspects of the age-gender association across datasets and modalities that extend beyond this characterization.

Regarding the use of the term “bias,” our original intention was to use it in a purely statistical sense (i.e., as a measurable ‘statistical bias’), as explicitly defined in the Guilbeault et al. (2024) *Nature* study on which our paper builds. This approach builds upon classic definitions of stereotypes as a form of statistical generalization, as in Allport’s canonical cognitively oriented definition of stereotypes as “exaggerated beliefs associated with a social category” that makes no direct claims about the evaluative nature of these exaggerated beliefs (1972 [1954], p. 191). However, we recognize that in the broader stereotype literature, “bias” can also carry a negative and evaluative connotation implying discriminatory judgment. In our revised introduction (copied at length in one of our replies to R1 above), we explicitly identify our focus on bias in this statistical sense.

We have added further revisions to our discussion (copied below) which highlight that our statistical findings do not directly identify the evaluative normative implications, and to the extent these can be gleaned from our findings, the patterns flag possible disadvantages for both women and men. On the one hand, our results indicate potential disadvantages for women by showing that ChatGPT assigns women fewer years relevant experience (holding occupation constant), while giving the highest scores of resume quality to older men. Additionally, new analyses, incorporated in response to other reviewers' recommendations, explore the relationship between status, prestige, and gender-age associations. These analyses indicate that for high-status and high-income occupations, online images are more likely to depict men as older than women. On the other hand, we find evidence that our participants report preferring to hire women (as indicated by their likert responses, Fig. S18); and excellent work from Ashley Martin and colleagues (as R1 points out) shows that, in some contexts, older men face more discriminatory judgment than older women.

In light of these complexities, we have refined our manuscript to clarify that our use of bias is intended purely in a statistical sense in the association of women with younger and men with older ages, rather than an inherent disadvantage for women. Our revised discussion emphasizes that our results do not directly indicate whether women or men are disadvantaged by the age-gender associations we observe. We greatly appreciate your feedback and have taken steps to ensure that our framing more accurately reflects the nuances of our findings. We copy these revisions here for your direct review.

From this discussion on pg. 17:

“How might the mass distortion of age-related gender associations online negatively impact women and men? Our results highlight several key ways in which older women are likely to be disadvantaged by this bias. For example, when generating resumes, ChatGPT not only assumes that women are younger, but also that they have less overall experience; consequently, ChatGPT is biased toward giving lower scores to resumes from younger women compared to older women, while giving the highest scores to older men. Yet, ChatGPT also gives higher scores to resumes from young women compared to young men, suggesting that young men may also be disadvantaged by this dual bias, a finding also supported by some of our experimental measures of participants' hiring preferences (Fig. S18). However, a selection bias favoring younger women and older men may further reinforce gender inequalities at the systemic level, whereby women are preferentially hired into roles with lower status and authority but denied mobility, while older men continue to enjoy their positions at the top. This resonates with our finding that online content is most likely to depict men as older than women for occupations with higher social status and wealth. Future research is needed to explore the ramifications of this dual bias on structural inequalities, especially in light of the sheer extent to which this bias permeates digital culture and mainstream algorithms.”

Reference:

Allport, G. W. (1979). The nature of prejudice (25th anniversary ed.). Cambridge, MA: Perseus Books.

Reply to Reviewer 2 (R2)

This is a really excellent manuscript. This research is very robust and the combination of methods is brilliant. This work not only evidences the case for the intersection of ageism and sexism, its a model of how to clearly and transparently draw together new AI methods and more traditional psychological experimentation to examine how the accumulation of biases (whether by human or artificial raters) reproduces structural cultural biases. I have no recommendations for this paper.

Peter Hegarty

- We are deeply honored to receive this generous endorsement and recognition of our work. We hope that the revisions inspired by this review process have further strengthened R2's appreciation for our work and its contribution.

Reply to Reviewer 3 (R3)

This manuscript provides extensive and convincing empirical evidence on the underrepresentation of older women' in the online sphere. The authors perform an impressive exercise across a range of textual and image datasets, showing that women are consistently represented as younger online compared to men. In addition, they show how generative LLMs replicate these age- and gendered patterns once asked to generate CVs. Finally, authors show, in a set of neatly designed experiments, that the bias in gender and age representations in Google images search results amplifies biases in participants' beliefs about ages and genders of people in different occupations.

This is without a doubt a very admirable undertaking, and I highly commend how thorough and detailed the authors' analyses are. The manuscript presents compelling evidence of biases in a world that increasingly relies on image-based content and generative LLMs. As such, it is valuable to the general scientific community and the public alike. I also commend the authors for sharing their underlying data and well-documented analysis code.

In addition, I am glad to see a thorough, large-scale test of concepts (i.e., gendered ageism) we consider well-established in social sciences, but only have scattered evidence of. I found the empirical evidence thorough and rather convincing.

- Thank you very much for your generous and thoughtful feedback! We deeply appreciate your kind words and are especially grateful for your recognition of the scope and rigor of our study.

That being said, I am listing some notes and questions:

A. I think this manuscript delivers two important messages: (1) There are inequalities in reality that are reflected in the online sphere (shown by the image and LLM analyses); and (2) our perception of social reality is further distorted by the inequalities reflected in the online sphere (shown by the experiment).

*Potentially, there is also an intermediate message: (1b) the inequalities in the online sphere are more severe than those in the society, which is once again amplified by (2). Authors suggest that “such age-based gender bias might systematically distort underlying sociodemographic realities”. Yet, we are not told to what extent the biases authors identify mirror or deviate from the social realities. If most of the results mirror the social reality of gender composition of different occupations (e.g., the ones from the WordNet categories used in Google image searches and GPT-2 evaluations), the interpretation of most of these findings is somewhat different than if biases present in the society are **also** amplified in the online sphere before biasing individual decisions. I would strongly encourage authors to be explicit about this.*

- We thank R3 for this excellent crystallization of the intuitions underlying our work and for astutely highlighting promising opportunities for further clarity and synthesis. As you will see in the replies to follow, your comments spurred us to undertake additional, extensive analyses benchmarking our observational data to underlying sociodemographic realities as captured by the U.S. census, revealing surprisingly clear evidence that, as you intuited, the inequalities in the online sphere are significantly more severe than those in society. With these new, clarifying findings in hand, we gained a deeper appreciation of both the significance of our Google Image search experiment and our resume audit study of GPT; namely, both identify ways in which mainstream algorithms amplify age-related gender bias in how we interface with representations of the social world in the online sphere. These developments led us to develop an improved framing that emphasizes what we view to be the core innovation and novel theoretical contribution of our paper. By examining online age-gender associations at an unprecedentedly large-scale across modalities and online platforms, we are able to produce a meaningful analysis of the extent to which these associations distort the underlying ground truth of age-gender associations in society, e.g. in occupations and industries. We copy these revisions in response to other replies from R3 below. We have also added extensive new data and analyses (further described in response to other replies), which shows that the intensity of age-related gender bias online systematically increases for occupations associated with high social status and wealth (and most of all for occupations in which the pay gap between genders is highest according to the census). Lastly, we show the extent to which these biases are reinforced and propagated through numerous widely used algorithms, including search engines and LLMs.

As our revised introduction emphasizes, these findings are of broad theoretical interest to the ongoing debate in social psychology and sociology about the accuracy of widespread stereotypes. Specifically, it addresses whether popular stereotypes capture true correlations in the characteristics of social groups, or whether they are systematically distorted by social and algorithmic processes. Age-gender associations

provide a uniquely apt context to pursue this question because of the ability to anchor age-related information in ground truth data about the biological age of men and women in society.

As R3 notes, prior work on gendered ageism has been scattered, limited in scale, and mostly based on survey studies or qualitative interviews in specific organizational and industry contexts. The inconsistencies in findings across these studies have raised the question of whether gendered ageism is indeed a culture-wide bias, or instead, an organization-specific problem. Our findings provide unprecedentedly clear evidence not only that age-related gender bias is indeed a culture-wide problem (permeating online images, videos, and massive, diverse corpora of online text), but also that this bias is a mass distortion of underlying sociodemographic realities that is systematically reinforced by some of the most influential algorithms in daily use today.

We believe this revised framing better highlights both the broad theoretical and empirical importance of our findings, not only in relation to gender inequality, but also in the study of stereotype formation, algorithmic bias, and ultimately the social construction of reality. We are indebted to R3 for the excellent suggestions, which have significantly deepened our contribution. Thank you in advance for taking the time to read through the introduction and discussion sections of our revised manuscript.

For your direct review, we copy our revised framing here. The results of our new analyses are copied in reply to other comments from R3.

Corresponding revisions to the introduction (starting on pg. 1):

“While few deny that stereotypes – generalizations about social groups^{15–17} – are harmful, a central question remains contested in social science: are common stereotypes generally accurate^{1–4} or socially distorted?^{5–8} Some argue that commonplace stereotypes accurately capture observable aspects of social groups, otherwise they would not gain such widespread adoption^{1–4,18,19}. Yet, others argue that most stereotypes are exaggerated or illusory^{5–8}. Assessing stereotype accuracy is challenging because stereotypes involve not only statistical associations (e.g., expected correlations among the features of a social group) but also normative judgments (e.g., that one group is superior to another) for which there is no well-defined ground truth^{15–17}. Even for statistical associations, identifying ground truth is difficult. In some cases, this stems from disagreement on how to measure the ground truth, such as enduring debates over how to measure intelligence²⁰ (a heavily stereotyped characteristic^{21,22}). Yet, even when there is agreement on the relevant constructs, there is often a lack of large-scale, quantifiable cultural data for measuring pervasive stereotypical associations and comparing these to ground truth indicators. As a result, research on stereotypes often yields inconsistent findings, calling into

question the pervasiveness and impact of these biases. In this study, we overcome these limitations in the analysis of age-related gender bias.

On the one hand, evidence abounds that older women face a dual bias at the intersection of gender and age. Policy reports^{11,12,23,24}, media coverage²⁵, and workplace interviews^{9,10,26} indicate that older women are discriminated against in hiring and promotion across industries (known as “gendered ageism”^{9,10,26,27}). This is related to a general statistical bias toward associating women with expectations of youth. From entertainment media to the workplace, women face persistent pressures to appear young, imposing a “beauty tax” with sizeable financial and time costs^{25,28,29}. This bias also manifests in everyday language. For instance, women in academia³⁰ and industry^{28,31–34} are much more likely than men to be referred to using infantilizing pronouns (e.g., ‘girls’), potentially undermining their perceived professional status. **These patterns suggest that age-related gender expectations may form a culture-wide statistical bias that influences how people view social roles and relationships across domains, echoing theories of gender as a central organizing frame in human perceptions of the social world and beyond**^{36,37}.

Yet, on the other hand, the statistical association between women and youth contradicts observable socioeconomic realities. Since the 1960s, women have consistently outlived men in the U.S. by as much as eight years, a gap that has been increasing^{38,39}. Census data on occupations presents similarly puzzling trends (Fig. S1). Over the last decade, there has been no correlation between the fraction of women in an occupation and its median age according to the U.S. census (Extended Data Figure 1; Table S1). There are also no clear differences in the age distribution of women and men throughout the workforce (Fig. S2). In fact, from 2009 to today, employed men in the U.S. have been more likely to be under the age of 40 than women, with no statistical difference in the fraction of employed women and men above 40 (Table S2)⁴⁰. Moreover, recent surveys fail to observe gendered ageism in certain organizational settings and even suggest that older women may be less impacted by stereotypes than older men^{41–43}. These inconsistent findings resonate with broader critiques against claims of enduring gender inequality, such as research showing declines in gender stereotypes over the last century in online text^{44,45}, as well as recent studies showing that hiring across industries increasingly favors women^{46–48}. **This dissonant landscape raises the question of whether age-related gender bias is an organization or industry specific problem, or whether it is a culture-wide distortion that continues to reflect and contribute to systemic inequalities.**

We argue that this uncertainty is fueled by the (i) lack of culture-wide multimodal data on the associations between gender and age, and (ii) the lack of computational methodologies for comparing these associations to ground truth indicators. To date, there have only been a handful of studies examining age-gender associations in small-scale surveys and interviews with professional women^{9,10,14,41,42,49–51}, or in sparse, non-representative observational studies of particular industries, such as celebrities and athletes in entertainment media^{26,28,52–54}. Yet, failing to observe age-related gender bias in small samples of a few contexts does not indicate its lack of prominence on

a culture-wide scale. Recent work in human cultural evolution finds that social biases in how people categorize the world frequently emerge only at scale^{55–57} and can manifest as exaggerated or even illusory beliefs^{58,59}. This suggests the alternative view that skewed associations between gender and age can emerge as a large-scale statistical bias that distorts underlying socioeconomic realities, despite inconsistencies across small-scale samples and contexts.”

Authors do use actual census-based shares of men and women in different occupations in their robustness checks, but only to check for correlations with their alternative measures. I wondered why they did not take a step further to show the extent to which the online-content-based estimates reflect the reality of at least the US census (even though online content is not only curated based on content from the US and other English-speaking countries).

- This is a fantastic point. This spurred us to undertake analyses which we think have considerably enriched both the overall framing of our paper and our statistical results. Initially, we did not pursue this particular analysis because at the time we did not find any U.S. census data that provided a breakdown of gender and age by occupation. Fortunately, thanks to R3’s insightful comment, we conducted a thorough review of the U.S. Census Bureau’s data offerings, and we were pleased to discover that they recently released a breakdown of gender and age by industry, from 2019 to 2023. While this dataset is not as granular as would be ideal, since it only provides aggregate data at the industry level across five overarching industries – sales, services, natural resources and construction, production and transportation, and management – it still provides a valuable industry-level perspective over recent years.

The five broad industries align with the occupations from our Wordnet dictionary, which we can link to the exact industry labels using the Census Bureau’s categorizations. This allows us to estimate the relationship between gender and age at the industry level by averaging the age associations by gender in Google images across all occupations within a given industry (first averaged within each occupation, and then across occupations at the industry level). Conveniently, the age groupings in the census data are similar to the age categories our participants used when providing age judgments of faces, allowing us to calculate direct correlations between the age codings from these distinct sources. The full supplementary section is provided below for ease of review.

At a high level, we find: (i) that the age associations of each gender at the industry level are significantly and robustly correlated across Google images and the US Census (spanning 2019 to 2023); yet (ii) the age gap between women and men is significantly higher (in an absolute sense) in Google images compared to the census, and (iii) in Google images, this age gap at the industry level always indicates men as older than women, whereas the census shows no consistent bias toward men being older than women across industries and census years. We view this as further evidence that online content distorts age-related gender associations relative to underlying ground-truth

distributions, and in a direction that is consistent with (and reinforces) biases relating to gendered ageism (i.e., toward associating women with youth).

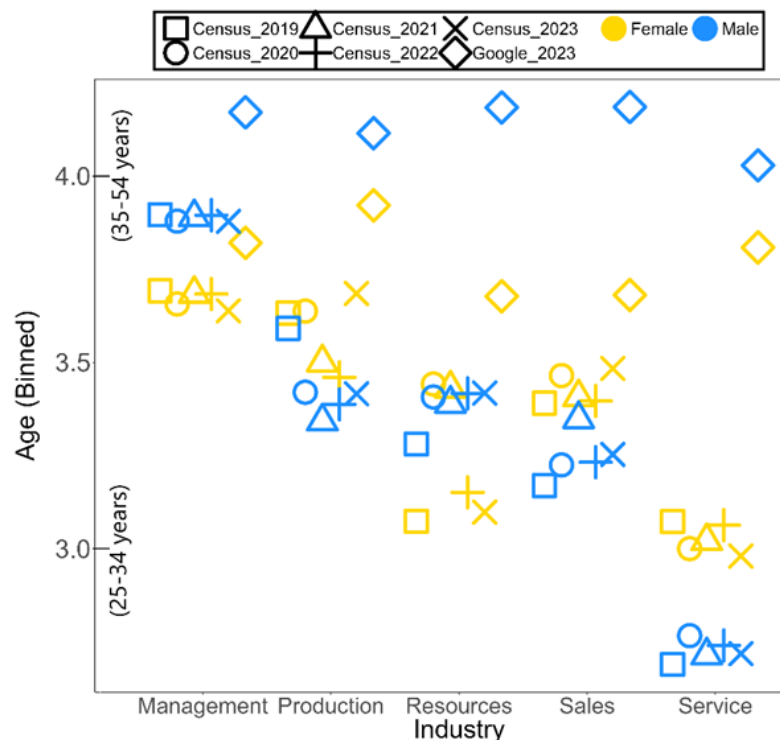
We have incorporated these findings into the manuscript through Extended Data Figures 2 and 3. We copy these revisions below for your review. R3 may also be interested in the additional analyses of the U.S. census data provided by Extended Data Figure 1 and via supplementary materials further indicating that gender-age associations in online media are distorted relevant to the ground truth. These additional analyses demonstrate a lack of correlation over the past decade between the percentage of women in an occupation and the median age associated with this occupation in the U.S. (Extended data figure 1; Fig. S1; Table S1). These revisions are also provided below for your direct review. (Note, we have also added new analyses using large-scale survey data and ground truth census data to identify the role of socioeconomic status in predicting age-related gender bias online; these results are copied in response to other comments from R3).

Census analyses added to the results section (pg. 7):

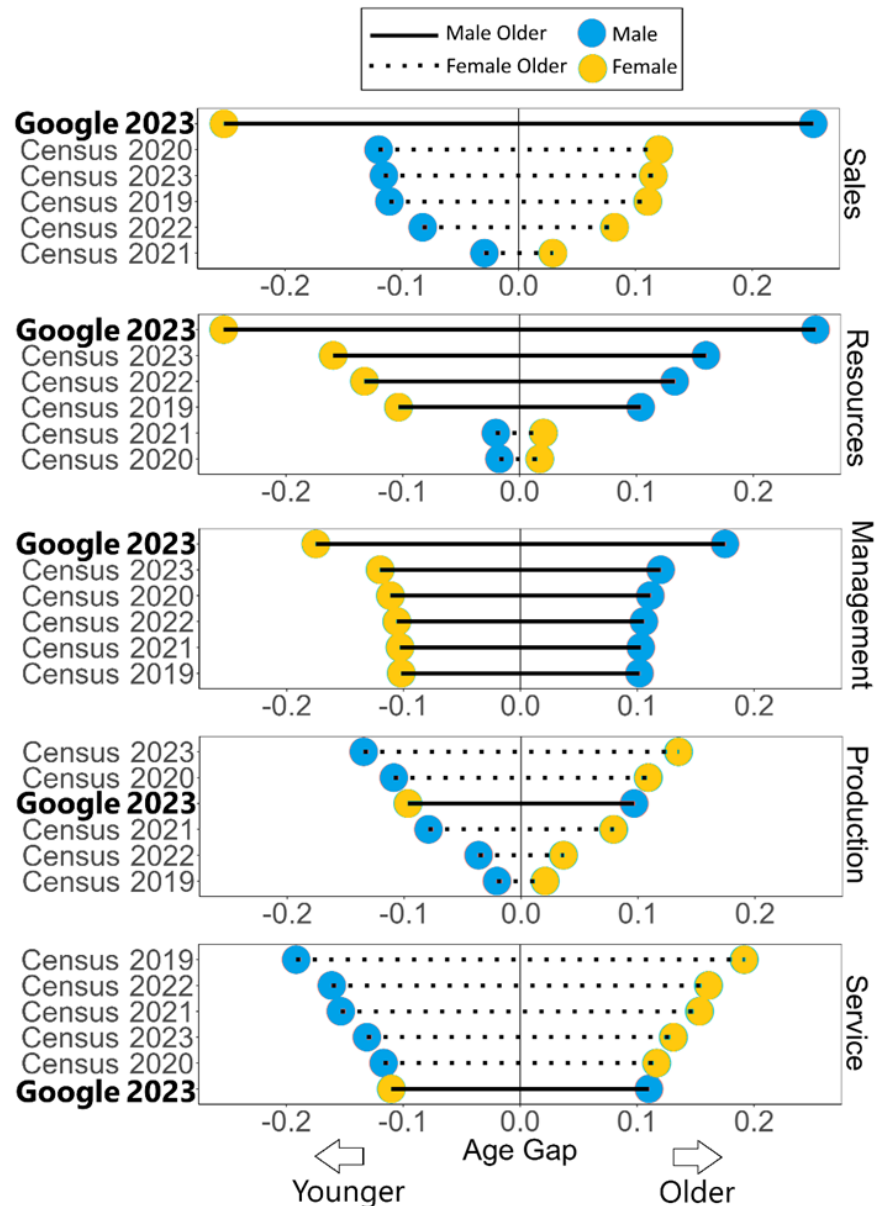
“Comparing to the Census

We now compare these findings to available ground truth data. We were able to match 867 social categories from our main Google image (Fig. 1A) dataset to occupational categories in the U.S. census. First, we observe that the average age rating for faces associated with occupations in Google is significantly correlated with the median age of people in these occupations according to the census (Tables S8 & S9; includes replications via Wikipedia). This indicates that the age estimates provided by the human coders are consistent with empirical age distributions across occupations. Yet, comparing to the census also shows how online images distort ground truth age-gender associations. The U.S. Bureau of Labor Statistics recently released a breakdown of the median age of each gender, from 2019 to 2023, across five industries: sales, services, natural resources and construction, production and transportation, and management. The census assigns each occupation to one of these industries, allowing those occupations matched in our Google image dataset to be assigned a census industry. We estimate the relationship between gender and age at the industry level by averaging the age associations in Google images across all occupations within a given industry (averaged within each occupation and then across occupations at the industry level). Conveniently, the census age groupings are highly similar to the age groupings the coders used when judging faces.

While gender-age associations in Google images and the census data are correlated at the industry level ($r = 0.13$, $CI = [0.11, 0.15]$, $p = 2.2 \times 10^{-16}$, Pearson Correlation, two-tailed; Extended Data Figure 2), Google images consistently display exaggerated and, in some cases, inverted trends that amplify the association between women and youth. Extended Data Figure 3 presents the absolute age gap between women and men by each industry, vertically ranked in terms of the magnitude of this gap, while also placing the older gender on the right side. In the industries of sales, resources, and management, Google images consistently present the highest age gap relative to all census years ($p < .001$ for all pairwise comparisons, Student t-test, two-tailed). Moreover, in each of these industries, Google images display men as older than women, while women are actually older than men for each of the census years examined in the sales industry, and for two of the years in the Resources industry. In the production and service industry, the magnitude of the age gap captured by Google images is not higher than all census years, yet the bias toward representing men as older is stable. In each census year examined, women are actually older than men in the production and service industry. It is only in Google images that men are older than women in these industries.”



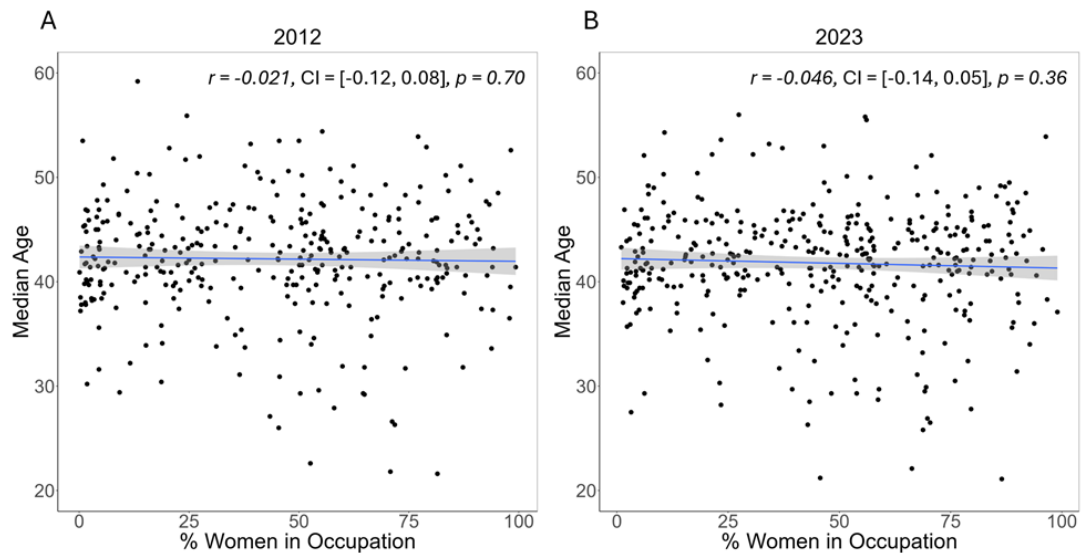
Extended Data Figure 2. Comparing the average age of women and men across industries in the U.S. Census (from 2019 to 2023) to the average perceived age of people in occupations from these same industries according to Google Images. The shape of the points indicates the data source, and the color of the points indicates the associated gender ($N = 867$ matched occupations).



Extended Data Figure 3. Comparing the magnitude and direction of the gender-age gap across industries in the U.S. Census (from 2019 to 2023) to the average perceived age of people in occupations from these same industries according to Google Images. Each panel shows the age gap between each gender for each industry separately. The midpoint of the age gap for each industry and data source is centered at 0 to help visually compare the magnitude of the age gap across datasets for each industry. Negative values along the horizontal axis indicate the gender that is associated with the lower age (relative to the midpoint), whereas positive values indicate the gender that is associated with the older age (relative to the midpoint). The color of the point indicates which gender falls on each side of the gender gap. Bold lines indicate cases where men are associated with a higher age than women for a given data source in a given industry; dotted lines indicate cases where women are older than men.

Census analyses added to the introduction (pg. 2):

“Yet, on the other hand, the statistical association between women and youth contradicts observable socioeconomic realities. Since the 1960s, women have consistently outlived men in the U.S. by as much as eight years, a gap that has been increasing^{38,39}. Census data on occupations presents similarly puzzling trends (Fig. S1). Over the last decade, there has been no correlation between the fraction of women in an occupation and its median age according to the U.S. census (Extended Data Figure 1; Table S1). There are also no clear differences in the age distribution of women and men throughout the workforce (Fig. S2). In fact, from 2009 to today, employed men in the U.S. have been more likely to be under the age of 40 than women, with no statistical difference in the fraction of employed women and men above 40 (Table S2)⁴⁰. Moreover, recent surveys fail to observe gendered ageism in certain organizational settings and even suggest that older women may be less impacted by stereotypes than older men^{41–43}. These inconsistent findings resonate with broader critiques against claims of enduring gender inequality, such as research showing declines in gender stereotypes over the last century in online text^{44,45}, as well as recent studies showing that hiring across industries increasingly favors women^{46–48}. This dissonant landscape raises the question of whether age-related gender bias is an organization or industry specific problem, or whether it is a culture-wide distortion that continues to reflect and contribute to systemic inequalities.”



Extended Data Figure 1. The correlation between the percentage of women in an occupation and the median age of people employed in this occupation according to the U.S. Bureau of Labor Statistics. Panel (A) shows the raw data (with each data point showing a single occupation) for 2012 (the correlation is non-significant; $r = -0.021$, $CI = [-0.12, 0.08]$, $p = 0.70$, Pearson Correlation, two-tailed, $n = 536$ occupations). Panel (B) shows the raw data (with each data point showing a single occupation) for 2023 (the correlation is non-significant; $r = -0.046$, $CI = [-0.14, 0.05]$, $p = 0.36$, Pearson

Correlation, two-tailed, $n = 594$ occupations). For all census years for which this data is provided in this format (from 2011 to 2023), there is not a single year with a statistically significant correlation between the fraction of women in an occupation and its associated median age (Table S1). Error bands show 95% confidence intervals.

I found that the current organization of the paper does not emphasize these two arguments; I think making them more explicit could benefit the paper. Especially the placing of the experimental results – while obviously connected to the analyses of bias in images – created a discontinuity between these two arguments. Rearranging the results could improve the flow of the paper and make the relationship between the two messages of the paper more salient.

- We completely agree with this suggestion. We have revised our results presentation to align with our improved introductory framing, which first asks (1) whether age-related gender bias is a culture-wide pattern in multimodal data, then examines (2) whether age-related gender bias in the online sphere reflects or distorts age-related gender biases in underlying sociodemographic realities (as captured by U.S. census on occupations and industries), and (3) assesses whether these biases are amplified through popularly used algorithms. To align with this narrative arc, our revised results section is structured in accord with the following trajectory. This involves the addition of new subsection headings in the results section to guide the reader. We indicate these headings in bold below as we outline the reorganization of our results section.

Age-Gender Distortions in Visual Content: First we show that age-related gender bias (skewing toward representing women as younger and men as older) pervades large-scale datasets of online images and videos from across popular platforms. **Comparing to the Census:** We then show how these visual biases systematically distort ground truth sociodemographic realities. Specifically, we benchmark our image data on occupations to census data on the distribution of women and men across industries, and find clear evidence that online images significantly exaggerate the representation of women (men) as younger (older), and in some cases, completely distort the underlying age distribution by showing women as younger than men in industries where the exact opposite is true according to the census. This is consistent with the claim that the association between women (men) with younger (older) ages is indeed a culture-wide bias that distorts sociodemographic reality. **Relationship to Social Status:** We further show how this bias is intensified as a function of the socioeconomic status of occupations, as captured by a nationally representative sample of people's perceptions, as well as by ground truth indicators of occupational prestige and yearly earnings provided by the U.S. Bureau of Labor Statistics.

Age-Gender Distortions in Online Text: Going further, the findings above raise the question of whether this is driven by the visual modalities specifically (e.g., through the use of image filters and cosmetics) or whether this bias truly pervades cultural thought through other pervasive modalities. This is why, secondly, we present our analyses of large-scale associations between women (men) and younger (older) ages in co-

occurrence patterns among concepts in massive bodies of text from across the internet as captured in nine distinct language models trained on billions of words. These findings strengthen our argument that age-related gender bias is a culture-wide statistical distortion of how gender is conceptualized on a mass scale across modalities, including images, videos, and text.

Algorithmic Amplification: Having documented age-related gender bias across modalities throughout the internet, our study is uniquely poised to theorize and empirically examine the critical role that online algorithms play in reinforcing this bias. To examine this possibility, we transition into the results of our experimental analyses of two of the most popular mainstream algorithms in popular use – i.e., the Google Image search engine and GPT. **Amplification via Google Search:** We begin by showing how using Google to search for images of occupations amplifies people’s biases toward misperceiving women as younger than men in the same occupations, and toward preferring to hire older men to older women. This demonstrates clear evidence of algorithmic amplification via Google Image search with measurable impacts on the psychology of internet users. **Amplification via Workplace AI:** We conclude by testing for algorithmic amplification more directly linked to the textual modality through an audit experiment examining age-related gender bias in GPT’s generation and evaluation of text-based resumes. This audit provides clear evidence that GPT assumes the female applicants are younger with less experience compared to male applicants to the same occupations, and more concerning, that the resumes associated with women are lower quality and less hireable compared to resumes from older men. This exposes another pathway through which mainstream algorithms amplify age-related gender bias. This is relevant beyond age-related gender bias as a study of how algorithms increasingly mediate and (in this case) distort our collective representations of the social world.

We believe this revised structure has significantly improved the flow and clarity of our results presentation, while also better highlighting their significance. All of our original analyses remain otherwise unchanged and are simply re-ordered with refinements and additions to the text to clarify the transitions between analyses. Our new analyses relating to ground truth census data and status effects are integrated and framed as complimentary to the main analyses of our original submission.

B. Whereas it is clear that the authors wanted to provide a comprehensive overview across various occupations/categories, I found the absence of any discussion of gender/age biases in relation to occupational status/prestige somewhat surprising. We know from the literature that women face a “glass ceiling” when it comes to entering prominent high-status positions (1). Such positions are occupied by older people; as women do not enter these positions, these older people are consequentially men. We also know that underrepresentation of certain groups in positions of power (2) or domains deemed to be of public and journalistic interest (3, 4) then leads to the these groups’ underrepresentation in the public sphere. I would be surprised if text and images online (which both largely build on past and current media content) would show different regularities.

Thus, the online content is not necessarily encoding (only) the age-based bias against women, but the broader structural inequalities that cause women's absence from high-status positions – and, consequentially, online image and text content. Authors treat all occupations/categories in their datasets as equal, but I see this as a missed opportunity to reveal something more profound about the structural mechanisms behind the inequalities they observe.

Thus, I think that the manuscript captures the gender-age nexus (e.g., secretaries are younger women, chairmen of the board are older men), but conceals how much of the bias is actually driven by the gender-status inequality. E.g., figure 3 highlights “cook” as associated with young women; but I would assume that “chef”, a higher-status “equivalent”, would be associated with more senior men. Then, the effects of gendered ageism (which seems to be the emphasis of the manuscript) and structural gender discrimination (which results in gender inequalities at higher-status levels, but is not necessarily directly a result of ageism per se) are confounded. This might not be of utmost importance in making the general point this paper is trying to make. Yet, providing a more solid understanding of potential mechanisms underlying the empirical findings would, in my opinion, make the conclusions more comprehensive and convincing.

- This insightful comment spurred us to undertake extensive analyses of how status and occupational prestige relate to the patterns of age-related gender bias in our large-scale Google image data depicting occupations. We undertook several ways of measuring the status of occupations: using a nationally representative survey of human ratings as well as official prestige scores and earnings data from the U.S. Bureau of Labor Statistics. Specifically, we collected new data (using a nationally representative survey with 1,002 participants) to evaluate the status and prestige of 867 occupations matched between our main Google Image data and the U.S. census from 2015 to 2022. This data presents clear evidence that occupational status predicts the intensity of age-related gender bias in online content consistent with the sociological literature as highlighted by R3. We now present these findings in Extended Data Figure 4.

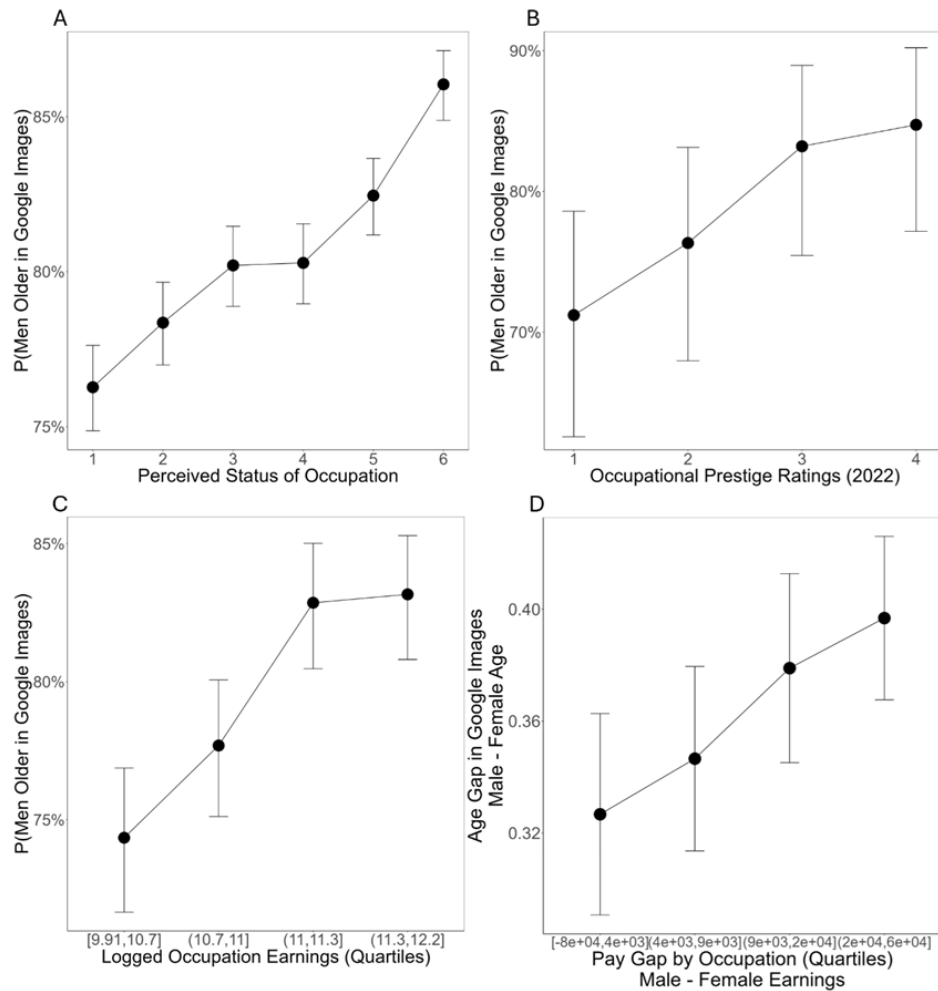
Here, we provide this figure and its associated prose in the manuscript directly for your review. Please note that we have also added a supplementary section, which demonstrates the robustness of these findings across a wide range of statistical controls, including controlling by participant and occupation. Please see section “A.1.12 Robustness of Status Associations” for extensive analyses, specifically Figure S9 and Tables S8-10. We believe these additional analyses have added considerable sociological depth to our findings that broaden its implications, not only for research on gender stereotypes, but also for understanding the role of status in shaping large-scale patterns of social categorization and evaluation that are propagated by mainstream algorithms. We sincerely appreciate your feedback and suggestions, which have meaningfully strengthened the manuscript.

New analyses added to the results section (pg. 7):

“Relationship to Social Status

Given the observational nature of these analyses, it is challenging to evaluate specific mechanisms driving these large-scale age-gender associations. Nevertheless, numerous patterns in our data are relevant to considering possible explanatory factors. One such consideration pertains to the hypothesis that gender stereotypes are most salient in occupations associated with higher status and prestigious, since high-status occupations often receive the most collective attention and praise, thus playing a prominent role in reinforcing gender expectations and norms of desirability^{80–82}. In a follow-up study, we recruited a nationally representative sample of U.S.-based Prolific users ($N = 1,002$) to evaluate the status and prestige of 867 occupations matched between our main Google Image data (Fig. 1A) and the U.S. census from 2015 to 2022. We find that occupations rated as higher status are more likely to elicit Google images in which men are older than women (Extended Data Figure 4A; $r = 0.08$, $t = 11.28$, $p = 2.2 \times 10^{-16}$, Pearson Correlation, Two-Tailed, $N = 867$ Occupations). Examples of occupations in the lowest or highest 5% of perceived social status are provided at the bottom of Extended Data Figure 4.

We then go beyond people’s subjective perceptions of status (which can widely vary⁸³) by using ground truth indicators. First, we reproduce this correlation using the U.S. Bureau of Labor Statistics’ measure of occupational prestige⁸⁴ (Extended Data Figure 4B; $r = 0.11$, $t = 2.5$, $p = .01$, Pearson Correlation, Two-Tailed, $N = 532$ Occupations). Next, we test whether the median yearly earnings associated with occupations according to the U.S. census also predicts age-gender associations in Google images. Indeed, the probability of men appearing as older in Google images is significantly higher for occupations with higher earnings (Extended Data Figure 4C; $r = 0.11$, $t = 7.39$, $p = 1.07 \times 10^{-13}$, Pearson Correlation, Two-Tailed, $N = 4,444$ pairwise comparisons at the census-year level from 2015 to 2022; yearly earnings logged). We even find that the gender pay gap^{12,24} — the extent to which men earn more than women in the same occupation — is associated with the extent to which men appear older than women in Google images, i.e., with the digital age gap (Extended Data Figure 2D; $r = 0.04$, $t = 7.305$, $p = .002$, Pearson Correlation, Two-Tailed, $N = 4,444$ pairwise comparisons at the census-year level from 2015 to 2022; yearly earnings logged). These results are robust to a range of statistical controls (Figs. S11 & S12; Tables S10-13), and resonate with longstanding concerns regarding connections between economic disparities and disparities in how genders are perceived and evaluated in the workplace^{12,24}.



Examples Based on Composite Measure (Panel A)

Low Perceived Status

janitor, meat packer, bellhop, salesclerk,
coal miner, window cleaner

High Perceived Status

astronaut, CEO, surgeon general, movie
star, prime minister, neuroscientist

Extended Data Figure 4. The age gap for occupations in Google Images is predicted by the perceived status of occupations, as well as by the median yearly earnings of occupations and the gender pay gap by occupation according to U.S. Census data from 2015 to 2022. Google image data is from Guilbeault et al. (2024; see Fig. 1A) and is based on 866 social categories matched to occupations in the U.S. census. We recruited a nationally representative sample of U.S.-based Prolific users ($N = 1,002$) to evaluate the status and prestige of 827 occupations matched between our main Google Image data (Fig. 1A) and the U.S. census. (A) The correlation between the perceived status of an occupation and the probability that men appear to be older than women in Google images of the occupation (status perceptions are averaged across a nationally representative sample of U.S. participants, $n = 1,002$ participants; an average of 27 participants rated each of occupations; data shown in six evenly spaced bins). See “Materials and Methods” for data collection and aggregation details.

Examples of occupations in the lowest (highest) 5% of perceived social status according to this measure are provided at the bottom of the figure. (B) The correlation between the U.S. Bureau of Labor Statistics' measures of occupational prestige (shown in quartiles) and the probability that men appear to be older than women in Google images of the occupation (532 occupations could be matched). (C) The logged median yearly earnings for an occupation (shown in quartiles) predict the probability that men appear to be older than women in Google images of the occupation. (D) The pay gap in median earnings for an occupation by gender (shown in quartiles) predicts the age gap in perceived age between men and women in Google images of the occupation. For (B) and (C), data are shown for the 753 occupations that could be associated with yearly earnings across Census years, 2015 to the present. Error bars show 95% confidence intervals.

Corresponding additions to the "Materials and Methods" in the manuscript (pg. 21):

"Collecting Judgments of Occupational Status"

We collected a nationally representative sample of 1,002 U.S.-based participants who provided their subject evaluations of the status and prestige of occupations. Each participant evaluated 20 randomly sampled occupations from a broader set of 867 Wordnet social categories that could be matched with corresponding occupations in the U.S. Census. Via randomization, each category was evaluated by 27 unique participants on average (minimum 15 participants). For each occupation, participants rated (i) its status using the following scale (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) and (ii) its prestige using the following scale (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). We also asked participants to rate the status/prestige through the standard question from the general social survey which asks participants to place occupations on a ladder containing 10 rungs, where the bottom rung indicates occupations with very low status, income, education, and prestige, whereas the highest run indicates occupations with very high status, income, education, and prestige (see Fig. S11). Participants' answers across all three questions were highly correlated (all paired Pearson correlations above 0.85; Fig. S9). In our main results shown in Extended Data Figure 4, we first average all participants' judgments of each occupation across the (i) status and (ii) prestige question, and then we assign each occupation a single status score by taking the mean of its average status and prestige score. In the appendix, we show that all of our results hold when examining each question separately, and when examining participants' judgments using the standard social status question from the General Social Survey (GSS) (Fig. S11; Tables S10-13). Note: Prolific's nationally representative sample of the U.S. population size allows for a maximum of 800 participants. However, this sample size was not large enough to gain sufficiently powered judgments across all 867 occupational categories, so an additional sample of U.S. participants was recruited

until all occupations reached a minimum of 15 evaluations from independent participants. All results are robust to a range of statistical controls (Tables S10-13)."

C. Related to Point B, having gone through the Google image search dataset, I had some questions/notes for the authors. I think the notes below also hold for the language model evaluations using WordNet categories.

Even when it comes to what authors denote as "ungendered" searches, some categories are clearly gendered. The most numerous categories related to women indicate youth by their nature (e.g., debutante, bachelorette, foster daughter, cheerleader, chorus girl, girl scout, check girl, working girl, brat, female child, princess, showgirl, valley girl). On the other hand, many of the categories for which many google images were evaluated for men clearly relate to very senior positions (e.g., cabinet minister, united states president, Mormon, government minister, civil rights leader, founding father, military volunteer, billionaire, civil leader, basketball coach, chief of state). Of course, the sole fact that so many women-related categories in WordNet are also youth-coded tells us a lot about the societal biases related to gender and age. However, one could not realistically expect to find older women online by searching for "chorus girls" or "cheerleaders". Are the authors confident that the age/gender gap they find does not largely stem from the selective nature of the words that were searched for in the Google dataset? This goes back to my point on the age/status nexus. Do results change if one drops words that are both female- and youth-coded? If they do – or do not – can this tell us something about the origin/nature of these gender/age biases?

On a more pedantic note, checking the age estimates for very gendered words (daughter, trophy wife, girl scout, half-sister), I saw that the results for these categories depicting men were evaluated as older than those depicting women. How meaningful are these averages? How much of an impact do they have on the final results? I suspect not much would change if such categories were excluded, but I found such observations rather difficult to grasp. Some other categories made rather little sense overall (e.g., ape-man or Capricorn).

- We thank R3 for this astute comment. Following this recommendation, we coded the age connotation of every category as either (i) directly connoting age (such as "youth" or "child"), indicated as 1, (ii) ambiguously connoting age (such as "amateur" or "beginner"), indicated as 0, and (iii) not explicitly connoting age (such as "player" or "friend") indicated as -1. To assess whether our main results are robust to the inclusion of age-coded categories, we replicated our analyses for both Google images and GPT2-Large embeddings after excluding all categories that either explicitly (1) or ambiguously (0) connote age. This resulted in dropping 393 categories, leaving 3034 categories for this analysis. We find that in both cases, our results are virtually identical in terms of effect size and statistical significance when excluding age-connoting categories. We copy this supplementary section below for ease of review. We have also added age connotation to our prior statistical models controlling for linguistic features in our analyses of both Google images and GPT2-Large associations. In both cases, the prior results presented

are essentially unchanged: the main effects reported are highly robust to including this control variable. For brevity, we do not copy these updated tables below.

We thus confirm that our results are not driven by categories connoting age. Given this robustness, we continue to present our main results using all categories in Wordnet to remain as comprehensive and agnostic as possible, while explicitly highlighting this key robustness test in the manuscript.

From pg. 16 of the supplementary appendix:

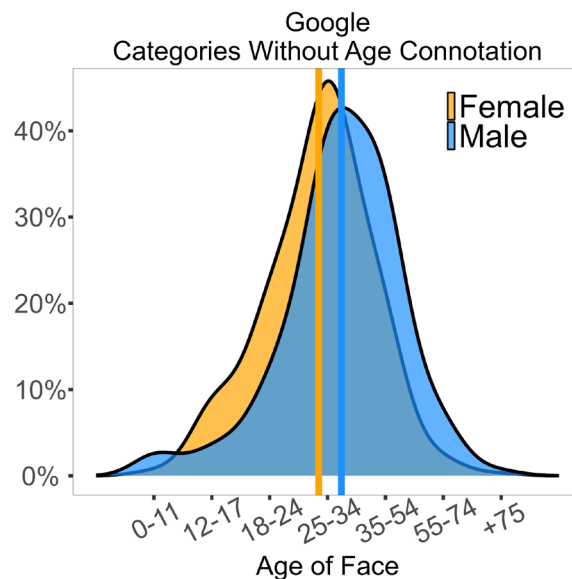


Figure S7. Women are represented as significantly younger than men in a large sample of images from Google associated with all social categories in Wordnet, while excluding categories that connote age, yielding 3,034 categories). Vertical lines indicate the average age for male and female faces in this dataset.

“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).”

From pg. 39 of the supplementary appendix:

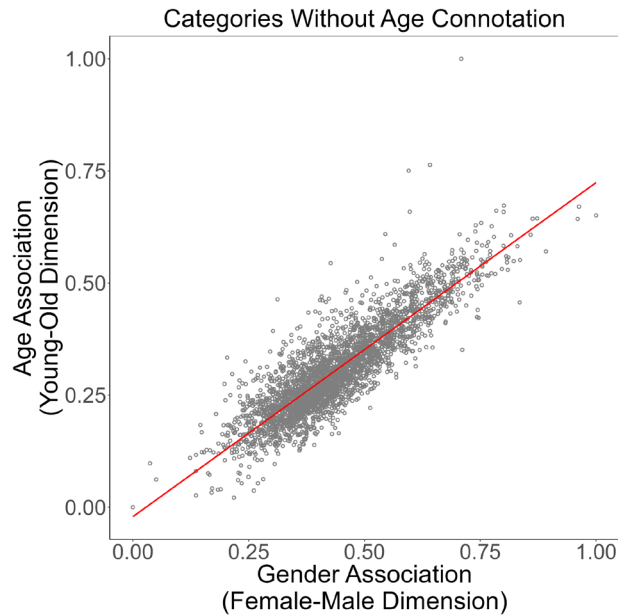


Figure S16. The correlation between age and gender associations in GPT2-Large for all social categories in Wordnet, while excluding categories that connote age, yielding 3,034 categories. The horizontal axis presents the gender association from 0 (female) to 1 (male), and the vertical axis presents the 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 S16). 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.”

D. I found the abundance and current presentation of different analyses and results somewhat overwhelming. By the time I reached “Measuring Bias in AI applications”, I had trouble recalling the first analysis, and had to step back to understand how this relates to all the other results. Perhaps the authors can find a way to ease the mental load of going through so many diverse, yet related results.

- We sincerely appreciate your perspective on the presentation of our analyses and the need for greater clarity in how they connect. Following your recommendations (as we detail in our reply above), we have carefully revised the framing, structure, and

transitions throughout the paper to create a smoother reading experience. We focused on adding clearer subsection divisions to better chunk out sections of results. For instance, our analyses that compare online images to the census are organized under a separate subsection entitled “comparing to the census”. We hope this study will better guide the reader and make our diverse yet related results more digestible and easier to organize into a broader, coherent whole.

I would also encourage the authors to make it clearer how the resume generation analysis complements and surpasses the results from the “Large-Scale Language Data” analysis. It is hardly surprising that, when asked to generate text based on their representations, LLMs reproduce the biases that are already contained in these representations. I would like a stronger emphasis on why we need both these analyses, presented separately, to gain a more complete image of the age-gender bias captured by the LLMs.

- Regarding the resume generation analysis, we appreciate your point about making its value more explicit. In our revised framing, we clarify that the intention of our experimental audit studies of both the Google Image search engine and GPT is to examine whether algorithms trained on large-scale internet data amplify age-related gender bias. We agree that this is a natural hypothesis that follows from our study’s unique achievement in documenting large-scale age-related gender biases in online images, videos, and text. In our view, its importance lies not in whether it is surprising *per se*, but rather in how it demonstrates the causal implications of the systematic distortion of age-related gender associations that our large-scale multimodal analyses of internet data capture. As our revisions clarify, the GPT analyses make a critical addition alongside our Google Image search experiment because the latter demonstrates algorithmic amplification in the visual modality, whereas the former demonstrates algorithmic amplification in the textual modality. Both modalities feature centrally in our large-scale analyses, and it is valuable to demonstrate algorithmic amplifications of age-related gender bias through each modality to ensure that such amplifications are not limited to a particular modality (e.g., as a result of image filters or cosmetics in visual presentations of women as compared to men). That said, our audit analyses of GPT are not as unsurprising as they may seem. Companies like OpenAI invest enormous sums of money into practices of curating quality datasets, fine-tuning machine learning processes, and imposing top-down filters to mitigate and eliminate social biases in the outputs of their algorithms. Given these efforts, it is not obvious or trivial that age-related gender bias would persist in GPT’s responses (let alone so sharply), especially in the consequential and common application of generating and evaluating resumes. On the whole, our study makes an important contribution not only by showing how age-related gender bias pervades multimodal data online and distorts underlying sociodemographic realities, but also by theorizing and demonstrating how this bias is amplified by the mainstream algorithms that are trained on this public internet data. As our revised discussion examines, this has far-reaching practical and theoretical implications. More broadly, these findings provide a clear demonstration of the increasingly central role that mainstream algorithms are playing in the social construction of reality by intensifying,

propagating, and entrenching distorted beliefs about social reality that emerge in the digital landscape. The exploration of age-related gender bias is a uniquely fitting stereotypical association for this inquiry, given the ability to isolate the statistical dimension of this association (via the correlation between gender and age), and the ability to compare multimodal online representations of age to ground truth indicators of the actual distributions of women and men of different ages throughout society.

To convey this reasoning, we have added the following clarifications to pg. 13 of the manuscript (directly relevant revisions in bold):

“These results show that exposure to online images through a popular search engine amplifies the distortion of age-gender associations in people’s beliefs. Some of this amplification may be due to image-specific presentation effects, such as the use of cosmetics and image filters to make women look younger than men, raising the question of whether the algorithmic amplification of age-related gender bias is limited to the visual modality. A more conservative test of algorithmic amplification is to test whether the implicit textual associations between age and gender in large language models can be traced to biases in the assumptions and judgments of these models in consequential settings. This is more conservative not only because textual associations between gender and age are often implicit (e.g., reflecting large-scale correlations in the co-occurrences of concepts), but also because popular LLMs are designed with a range of filters to avoid perpetuating social biases. Yet, given the only scattered recognition and demonstration of gendered ageism throughout prior research, it is possible that such models do not effectively filter for age-related gender bias. To test this, we conducted a systematic audit study of age-related gender bias in a workplace application of ChatGPT.”

E. Figure S9 in the supplement highlights a finding I find interesting, but which remains neglected as it is not fully aligned with the broader narrative: in the experimental results, young women are considered more hireable for the same positions compared to young men. Could it be that younger women are deemed more competent for some positions than young men? The situation then reverses for older people. I am once again wondering to what extent the job status – and the perception that there are no women in such high-status jobs – moderates this shift. That is, even though negative stereotypes surrounding women’s ageing prevail very broadly, and particularly in some industries (5), I wonder how much of the effect is driven by women’s absence from high-status categories.

References

1. D. A. Cotter, J. M. Hermsen, S. Ovadia, R. Vanneman, The Glass Ceiling Effect. Soc. Forces 80, 655–681 (2001).
2. E. Shor, A. Van De Rijt, A. Miltsov, V. Kulkarni, S. Skiena, A Paper Ceiling: Explaining the Persistent Underrepresentation of Women in Printed News. Am. Sociol. Rev. 80, 960–984 (2015).

3. P. England, S. Li, Desegregation Stalled: The Changing Gender Composition of College Majors, 1971-2002. *Gend. Soc.* 20, 657–677 (2006).
4. E. Shor, A. Van De Rijt, B. Fotouhi, A Large-Scale Test of Gender Bias in the Media. *Sociol. Sci.* 6, 526–550 (2019).
5. A. E. Lincoln, M. P. Allen, Double Jeopardy in Hollywood: Age and Gender in the Careers of Film Actors, 1926–1999. *Sociol. Forum* 19, 611–631 (2004).

- Once again, we find this connection to be fascinating and highly generative. We are deeply grateful for R3's suggestion to explore correlations with status and prestige in our age- and gender-related data. Your observations have helped us reflect more deeply on the role of occupational status and prestige in our results.

As we elaborate in detail above, our revised study now provides clear support for the prediction that age differences across genders are more pronounced for high-status occupations (as measured both by a nationally representative survey of people's perceptions of status and prestige, as well as by official prestige scores and yearly earnings data associated with occupations according to the U.S. Bureau of Labor Statistics from 2015 to 2022). In light of R3's additional comments here, we agree that these findings have important implications regarding our surprising finding that younger women are rated as more desirable as hires than younger men (now Fig. S18). We agree that this pattern warrants more direct attention in discussing our main results. However, we want to highlight that we are unable to establish a direct causal link between these patterns, due to the large-scale observational nature of our data, as well as the complex interdependencies between status perceptions and occupational stereotypes. In addition, given the already extensive scope of our analyses in need of streamlining (as R3 also highlights), we believe that properly exploring the causal dynamics underlying these patterns is best pursued in future research. To acknowledge this important insight, we have included the following note in our discussion highlighting this pattern and indicating these open questions as an exciting direction for future research. We are grateful for your valuable input in helping us refine our discussion. We have also included the key references recommended above into our revised manuscript.

From the discussion (pg. 17):

“How might the mass distortion of age-related gender associations online negatively impact women and men? Our results highlight several key ways in which older women are likely to be disadvantaged by this bias. For example, when generating resumes, ChatGPT not only assumes that women are younger, but also that they have less overall experience; consequently, ChatGPT is biased toward giving lower scores to resumes from younger women compared to older women, while giving the highest scores to older men. Yet, ChatGPT also gives higher scores to resumes from young women compared to young men, suggesting that young men may also be disadvantaged by this dual bias, a finding also supported by some of our experimental measures of participants' hiring preferences (Fig. S18). However, a selection bias

favoring younger women and older men may further reinforce gender inequalities at the systemic level, whereby women are preferentially hired into roles with lower status and authority but denied mobility, while older men continue to enjoy their positions at the top. This resonates with our finding that online content is most likely to depict men as older than women for occupations with higher social status and wealth. Future research is needed to explore the ramifications of this dual bias on structural inequalities, especially in light of the sheer extent to which this bias permeates digital culture and mainstream algorithms.”

Referee #3 (Remarks on code availability):

I have only ran a portion of the code; but I found it easy to read and follow. I also succeeded in producing several figures in the manuscript.

We are very happy to hear that our github was accessible and our code easy to run and validate in this manner. Thank you for putting in this generous effort.

The specific descriptions of individual variables and the details about the information contained in each dataset were less clear.

Thank you for this valuable feedback. To help address this issue, we have added an extensive read.me to the github page which provides an explicit description of the meaning of each column in our dataset to help improve interpretability. Please note that to align with our revised framing, we have updated the name of the github repository: https://github.com/drquilbe/mass_distortion_age_gender_online. All new analyses and data are publicly available at our updated project github.

Referee #4 (Remarks to the Author):

Summary:

Using large-scale image and text data the present work showed a persistent bias to represent women as younger than men online. Furthermore in an experiment, participants prompted to use a search engine for an occupation (vs. to search for control concepts) showed an enhanced female-young bias.

I found the work to be original and significant. I also found it to be generally methodologically sensible. I reviewed the main text, SI, and preregistration. Below I offer questions, comments, and suggestions.

- We are happy to hear that you appreciate the novelty and significance of our work. Your comments have helped to further enrich and deepen our contribution. Thank you in advance for your thoughtful and constructive recommendations.

Comments:

The authors anticipated a key concern I had, that is, that their results reflect bias in the annotators/coders and not bias in the images. I was reasonably convinced by their efforts to rule out this alternative account. One comment, for A. 1. 4. it seems to me that the key analysis is whether the relationship (slope) between actual age and perceived age varies by image gender (i.e., the actual age x image gender interaction term in a regression predicting perceived age). I could not find this clearly reported in the SI. What am I missing?

- We are grateful for this very helpful recommendation. We agree that this analysis lends further clarity and interpretability to our analysis. We have run your recommended model and can confirm that there is no significant interaction between the actual age and the perceived age of each validation face as a function of its gender. We have incorporated these analyses in our revised appendix. We have decided to add this model alongside our additional analyses because we believe that each model buttresses our argument in complementary ways. We copy these additions below for ease of review.

From pg. 12 of the supplementary appendix:

“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 ($\beta[\text{True Age X Male}] = 0.00$, $\text{CI} = [-0.01, 0.01]$, $p = 0.97$). In fact, the effect of this interaction is essential 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.”

(Table copied below)

Variables	Beta	95% CI ¹	p-value
True Age	0.91	0.90, 0.92	0.000000e+00
True Gender			
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+".

Why is the manuscript framed around the invisibility of older women rather than the invisibility of younger men? See especially Figure 4C. It appears resumes for younger women were scored higher than for younger men, based on the fitted line at least. I believe that there might be good reasons to focus on the invisibility of older women (e.g., Figure 2B, the experiment lowered women's age estimate but did not raise men's).

- R4's comment here echoes the recommendations of the other reviewers. We agree that our prior framing around the "Invisibility of Older Women" did not sufficiently capture the richness and depth of our findings. As R4 astutely highlights, while some of our findings were consistent with this framing, others were less so, such as the evidence from both our Google search experiment and our resume audit of GPT indicating a potential preference for hiring younger women over younger men.

In our revised introduction, we clarify our core aim as documenting statistical biases in the association of women and men with age using large-scale multimodal data, and then evaluating the extent to which these patterns online reflect or distort underlying sociodemographic patterns (e.g., in the U.S. census data on the true age of women and men in occupations and across industries). As our revised results illustrate, we find clear evidence not only that age-related gender bias is culture-wide (permeating online images, videos, and massive, diverse corpora of online text), but also that these biases exaggerate and distort the underlying age differences of men and women as measured in census data (see Extended Data Figures 1–3, copied below). We have also added extensive analyses indicating that the strength of this age-related gender distortion is significantly higher for occupations associated with higher prestige and wealth (please see Extended Data Figure 4 and the new results subsection in the manuscript entitled "Relationship to Social Status").

Our results regarding our Google image search experiment and resume audit of GPT are now more precisely framed as an examination of how these distorted online biases are amplified by mainstream algorithms that mediate how we interface with digital representations of the social world. For our purposes, we are agnostic to whether these distortions and their amplifications stem from a bias in favor of younger women or a bias against older women, or some combination of both.

Our revision highlights that our findings do not highlight a singular normative takeaway, since our evidence suggests ways in which women are both disadvantaged and possibly advantaged by these distortions as a function of their age. For instance, while we find that older women are rated as less hireable in both our experiment and by GPT's resumes, some of our measures indicate that younger women are rated as more hireable than younger men (though still less hireable than older men). In both cases, our broader argument holds that the observed statistical bias in the association between gender and images in large-scale multimodal data systematically distort underlying socio-demographic realities, and these biases (which are complex in nature) are nevertheless demonstrably amplified and reinforced by search engines and LLMs.

I would have liked to see more focus on effect sizes throughout. For instance, the age-gender association appears to be twice as large in the non-gendered searches than the gendered searches (p. 4). What do the authors make of this? How should readers think about the size of these effects? Any benchmarks?

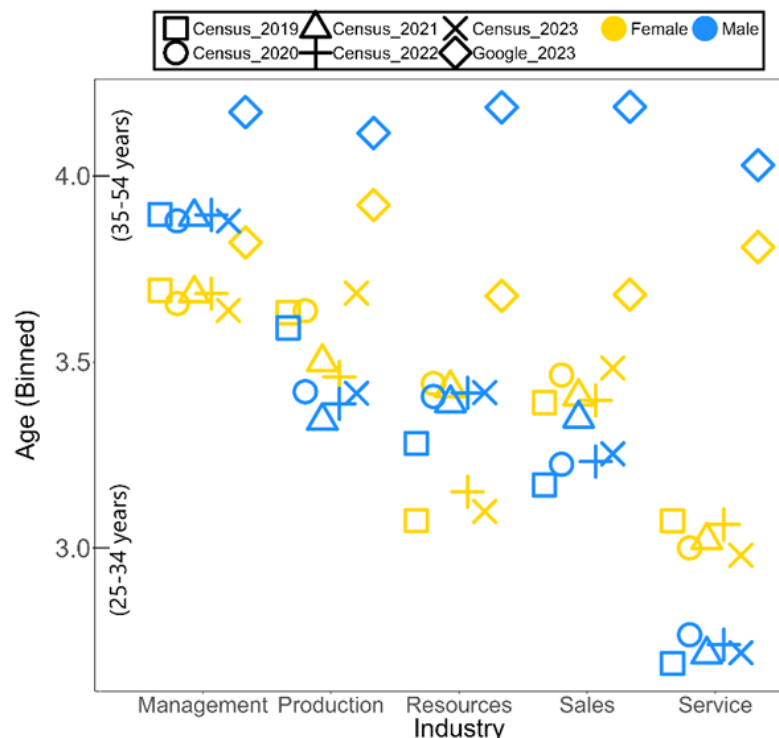
- Thank you for this helpful comment. We have included indicators of effect size (in terms of the difference in mean age across genders, which occasionally refers to the average ordinal age bin number) in the statistical reporting of all of our image results presented in Figure 1 (in addition to comparable effect sizes already provided in the prose presenting the results of our Google Image search experiment and our resume audit of ChatGPT). As well, we now much more thoroughly benchmark our image results against ground truth census data. We hope that these comparisons provide further useful context to help readers additionally appreciate the qualitative significance of these effect sizes. Please see Extended Data Figure 2 and 3, copied below for your review along with the manuscript prose that references these figures. Both figures visually display the effect size of the age gap observed among women and men, benchmarked against the ground truth age gap across genders in the U.S. Census across industries. We are happy to provide further information as requested.

From pg. 7 of the manuscript:

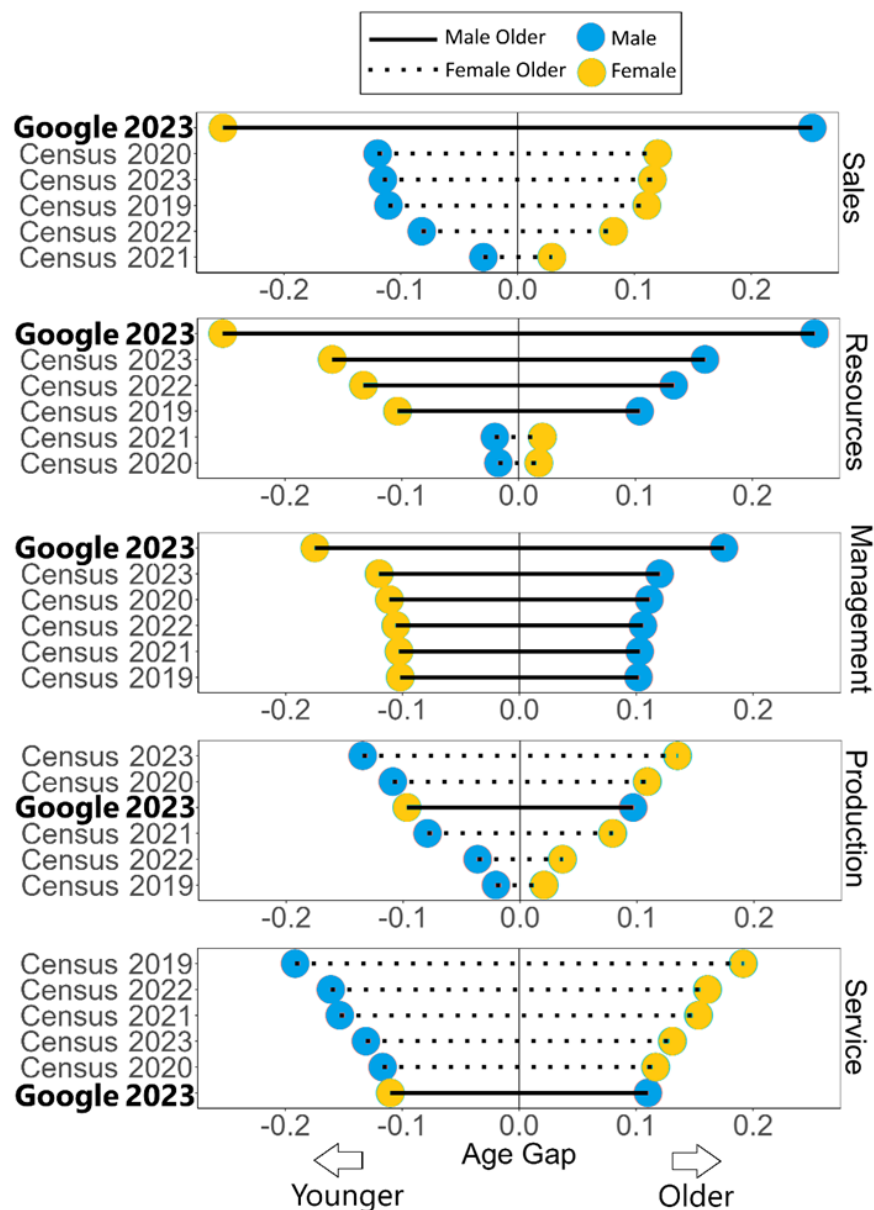
“Comparing to the Census

We now compare these findings to available ground truth data. We were able to match 867 social categories from our main Google image (Fig. 1A) dataset to occupational categories in the U.S. census. First, we observe that the average age rating for faces associated with occupations in Google is significantly correlated with the median age of people in these occupations according to the census (Tables S8 & S9; includes replications via Wikipedia). This indicates that the age estimates provided by the human coders are consistent with empirical age distributions across occupations. Yet, comparing to the census also shows how online images distort ground truth age-gender associations. The U.S. Bureau of Labor Statistics recently released a breakdown of the median age of each gender, from 2019 to 2023, across five industries: sales, services, natural resources and construction, production and transportation, and management. The census assigns each occupation to one of these industries, allowing those occupations matched in our Google image dataset to be assigned a census industry. We estimate the relationship between gender and age at the industry level by averaging the age associations in Google images across all occupations within a given industry (averaged within each occupation and then across occupations at the industry level). Conveniently, the census age groupings are highly similar to the age groupings the coders used when judging faces.

While gender-age associations in Google images and the census data are correlated at the industry level ($r = 0.13$, $CI = [0.11, 0.15]$, $p = 2.2 \times 10^{-16}$, Pearson Correlation, two-tailed; Extended Data Figure 2), Google images consistently display exaggerated and, in some cases, inverted trends that amplify the association between women and youth. Extended Data Figure 3 presents the absolute age gap between women and men by each industry, vertically ranked in terms of the magnitude of this gap, while also placing the older gender on the right side. In the industries of sales, resources, and management, Google images consistently present the highest age gap relative to all census years ($p < .001$ for all pairwise comparisons, Student t-test, two-tailed). Moreover, in each of these industries, Google images display men as older than women, while women are actually older than men for each of the census years examined in the sales industry, and for two of the years in the Resources industry. In the production and service industry, the magnitude of the age gap captured by Google images is not higher than all census years, yet the bias toward representing men as older is stable. In each census year examined, women are actually older than men in the production and service industry. It is only in Google images that men are older than women in these industries.”



Extended Data Figure 2. Comparing the average age of women and men across industries in the U.S. Census (from 2019 to 2023) to the average perceived age of people in occupations from these same industries according to Google Images. The shape of the points indicates the data source, and the color of the points indicates the associated gender ($N = 867$ matched occupations).



Extended Data Figure 3. Comparing the magnitude and direction of the gender-age gap across industries in the U.S. Census (from 2019 to 2023) to the average perceived age of people in occupations from these same industries according to Google Images. Each panel shows the age gap between each gender for each industry separately. The midpoint of the age gap for each industry and data source is centered at 0 to help visually compare the magnitude of the age gap across datasets for each industry. Negative values along the horizontal axis indicate the gender that is associated with the lower age (relative to the midpoint), whereas positive values indicate the gender that is associated with the older age (relative to the midpoint). The color of the point indicates which gender falls on each side of the gender gap. Bold lines indicate cases where men are associated with a higher age than women for a given data source in a given industry; dotted lines indicate cases where women are older than men.

I believe Reddit contributors have been documented to be extremely male-dominated. (Same for other text, Wiki, news) I understand that the results (e.g., p. 9) replicate across training algorithms and corpora. But how does the overrepresentation of men as contributors to the training set(s) influence the conclusions?

- This is an important point. R4 is correct that it is quite plausible that genders may differ in the extent to which they contribute to the textual data in particular platforms. The standard stance in the field of natural language processing is that by training models on incredibly large bodies of text, the embedding spaces that form capture the core meanings of words at the population level (i.e., their shared definitions) rather than the opinions or attitudes of specific groups. This argument has been made since the first static embedding models were popularized in the early 2010s, and it is held to be even more true with the recent GPT models that are trained on nearly the entire transcribable internet. This is different from cases where models are trained on small datasets for the purpose of understanding group-specific or context-specific biases in meaning, where these biases are typically measured by comparing embeddings extracted from smaller-scale settings to the more generalized embeddings formed from population-level language data. In our study, we are reassured by the strong replicability of our results across language models trained on distinct, large datasets collected at different times from different online platforms, which suggests robustness to different populations of data providers. It is further reassuring that our results replicate strongly using images and textual data from Wikipedia, since the data on this platform is shaped by community norms and institutionalized evaluation practices intended to neutralize language and promote factuality, despite known gender differences among Wikipedia editors. That said, it may be possible that the prevalence of age-related gender bias online is influenced by inequalities of gender representation in data producers. We agree this is an important topic for future research, which is why we have included the following sentence (in bold) in our revised discussion:

“[...] A fascinating question for future work is to explore the extent to which the aesthetic norms, fictionalized representations, and hiring biases of entertainment media spillover into the mass distortion of age-gender associations in other areas of social life. **A related question concerning supply-side factors concerns how age-related gender bias in popular algorithms may stem from inequalities in the gender of data contributors online; studies suggest that Reddit users⁹³ and Wikipedia editors⁹⁴ are disproportionately male, and textual data from these platforms are frequently mined for training AI models. Training AI on datasets with greater gender equality in data contributors may provide an effective mitigation strategy.**”

Given current preregistration standards, I suggest adding a table or similar to the SI to succinctly summarise adherence to and/or deviations from the preregistration. On my read, the

preregistration appeared to be largely in line with what was reported. But I encourage greater clarity on this for both hypotheses and analyses to make it easier for readers to assess.

- This is an excellent suggestion. We have added a table to the supplementary appendix that details each of the hypotheses we preregistered, while also indicating whether they were confirmed and the corresponding figure or table that presents the results of our preregistered analysis. As our discussion of this table emphasizes, all of our main preregistered hypotheses were supported using the analysis strategies we preregistered. We thank R4 for making this request because it allowed us to realize that we forgot to include the results associated with our “Main Hypothesis 3” in our preregistration. We have since conducted these analyses precisely following the preregistered analysis plan, and the results support our hypothesis. For exhaustiveness, we have included the results associated with this hypothesis in the appendix. These revisions to the SI are copied here for your direct review. In the manuscript, we point to this table at the end of our experimental results section.

From pg. 51 of the supplementary appendix:

“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>R²</i>	0.313		
<i>Adjusted R²</i>	0.272		
<i>No. Obs.</i>	4,829		
<i>df</i>	277		
<i>AIC</i>	29,533		
<i>Residual df</i>	4,551		
<i>Sigma</i>	5.01		
¹ CI = 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* ($\beta[\text{Male Uploaded Image}] = 0.72$, $\text{CI} = [0.35, 1.10]$, $p = .0001$); 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 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)."	Supported	Figure 2C; S17
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 preregistered 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 bias^{21–23}. We think that exploring the methodological and psychological drivers underlying these surprising patterns is an interesting direction for future work.”

Minor points:

- I'm not sure what “mere cultural aesthetic” means in sentence 1.

- We agree with R4 about the lack of clarity in this phrasing. We have removed it from the manuscript to avoid confusion.

- I found Figure 2.C unclear. Ultimately I was able to understand it, but I had to reread the relevant paragraph on p. 8 several times. For instance, ‘gender associations’ is unlabelled making it unclear what the poles are.

- We thank R4 for highlighting this important oversight. We have now added arrows and labels to Panel C of (now) Figure 3 to make explicit what the poles are, thereby improving the readability of this figure. The figures were re-ordered to improve the clarity of our results presentation and better align with our revised framing.

- Word embeddings capture mere co-occurrence as well as higher-order co-occurrence. That is, two words can be similar in vector space despite co-occurring with each other infrequently due to their frequent co-occurrence with the same other sets of words. I had trouble squaring this with the description on p. 9 of the gender dimension the authors constructed, which seemed to imply that it only captured mere co-occurrence.

- We agree with this point. Initially, we gave the example of words co-occurring in the same sentence to help readers without expertise in these models to get a sense of how they work. To make sure this pedagogical simplification does not puzzle those who are more familiar (or mislead those who are not), we have added the following clarification to the manuscript along the lines suggested by R4. We copy this revision here for your direct review (changed in bold):

“[...] we go beyond visual content by examining the extent to which age-related gender bias permeates large-scale patterns of conceptual associations and verbal expression in massive bodies of online textual data. For this purpose, we leverage word embedding models that construct a high-dimensional vector space based on the co-occurrence of words (e.g., whether two words appear in the same sentence), such that words with similar meanings are closer in this vector space. **Technically, these embedding spaces also capture higher-order similarities based on whether**

words co-occur in similar linguistic contexts (i.e., in association with related sets of words), without requiring words to directly appear together. ”

Reply to Reviewers

Referee #1 (Remarks to the Author):

I was Reviewer 1 on the previous version of this manuscript. My biggest concerns revolved around the theoretical framing of the paper, especially the argument around “older women invisibility” and the surprising lack of emphasis on what I thought was the more striking pattern in the data, which was that women of all ages seem to be perceived as act/look younger than comparably aged men.

I really applaud the authors’ attention to my critiques, and also like their reframe on comparing perceptions to ground truth vis-à-vis age-gender perceptions. I agree that age is a more “objective” metric that serves as a useful comparison between stereotypes and reality. Thus, I think the theoretical contribution of the paper is far clearer than it was previously. This reframe, compared with the additional analyses presented, convince me that this paper is worthy of publication.

My last remaining request is that the authors do a more explicit job of highlighting why the authors selected age as their “ground truth” objective criterion. The Abstract in my opinion does not make this clear, nor does the Intro. I will leave it up to the authors to decide how to do it, but in general, I actually think the authors made a stronger case in their point-by-point response about how age serves as a useful objective metric for comparing the accuracy of stereotypes. I would like to see this same level of care in the main text, as well.

Once again, I commend the authors on a convincing reframe, analysis, and important paper.

- We are delighted to receive this positive endorsement from R1. Thank you once again for your expert insight and recommendations, which have led to several

major improvements to the paper. Reviews like yours are a testament to the power of the scientific community in helping projects reach their fruition. We agree with your remaining recommendation to better highlight the key role of age as a uniquely clear ground truth anchor for evaluating the accuracy of corresponding gender-age stereotypes. We have updated the abstract and introduction accordingly. The revisions are provided below, in bold:

From the abstract:

‘Are widespread stereotypes accurate^{1–4} or socially distorted?^{5–8} This debate is limited by the lack of large-scale multimodal data on stereotypical associations and the inability to compare these to ground truth indicators. Here, we overcome these challenges in the analysis of age-related gender bias^{9–14}, **for which age provides an objective anchor to evaluate stereotype accuracy**. Despite no systematic age differences between women and men in the workforce, we find that women are represented as younger than men across occupations and social roles in nearly 1.4 million images and videos from Google, Wikipedia, IMDb, Flickr, and YouTube, as well as in nine language models trained on billions of words from the internet. This age gap is most stark for content depicting occupations with higher status and earnings. We then show how mainstream algorithms amplify this bias. A nationally representative, preregistered experiment ($N=459$) finds that Googling images of occupations amplifies age-related gender bias in participants’ beliefs and hiring preferences. Furthermore, when generating and evaluating resumes, ChatGPT assumes women are younger and less experienced, while rating older male applicants as higher quality. Our study shows how gender and age are jointly distorted throughout the internet and its mediating algorithms, revealing critical challenges and opportunities in the fight against inequality.”

From the introduction:

“While few deny that stereotypes – generalizations about social groups^{15–17} – are harmful, a fundamental question remains contested: are common stereotypes accurate^{1–4} or socially distorted?^{5–8} Some argue that commonplace stereotypes accurately capture observable aspects of social groups, otherwise they would not gain such widespread adoption^{1–4,18,19}. Yet, others argue that most stereotypes are exaggerated or illusory^{5–8}. Assessing stereotype accuracy is challenging because stereotypes involve not only statistical associations (e.g., expected correlations among the features of a social group) but also normative judgments (e.g., that one group is superior to another) for which there is no well-defined ground truth^{15–17}. Even for statistical associations, identifying ground truth is difficult. In some cases, this stems from disagreement on how to measure the ground truth, such as enduring debates over how to measure intelligence²⁰ (a heavily stereotyped characteristic^{21,22}). Yet, even when there is agreement on the relevant constructs, there is often a lack of large-scale, quantifiable cultural data for measuring stereotypical associations and comparing these to ground truth indicators. As a result, research on stereotypes

often yields inconsistent findings, calling into question the pervasiveness and impact of these biases. **In this study, we overcome these limitations in the analysis of age-related gender bias, which not only involves biological age as an objective anchor for evaluating stereotype accuracy, but which also can be linked to large-scale statistical biases in how the ages of women and men are depicted online.**"

Referee #3 (Remarks to the Author):

I appreciate the authors' thorough and thoughtful engagement with my and other reviewers' comments and suggestions.

I find the revised framing which now explicitly tackles the comparison between the ground truth and online bias, as well as the two types of bias amplification, more coherent and impactful. I was glad to hear that the authors found my suggestion on this comparison useful.

I do find the granularity of the US Census data less than ideal; at the same time, I am convinced that the authors provided insights on the ground truth of average ages per industry to the best of their abilities.

The additional analyses on status add depth to the discussion; I find the insight that the age gap persists despite controlling for different facts of status informative - both theoretically and empirically. I was impressed by the breadth and comprehensiveness of status-related analyses conducted by the authors.

These revisions comprehensively address my concerns, and I am convinced by the arguments in the authors' response to reviewers.

- We are deeply grateful for the depth and quality of R2's feedback on previous versions of this manuscript. Your insights have led to many significant improvements to our paper. We are honored to receive your positive endorsement for publication.

Referee #4 (Remarks to the Author):

I reviewed a previous version of this manuscript. I read the author's response to my comments as well as the new version of the manuscript. I found this revision to be responsive to my comments and those of the other reviewers. I think an already strong manuscript is now clearer with the interpretation of the data better articulated. I have no further comments.

- We are deeply grateful for the depth and quality of R4's feedback on previous versions of this manuscript. Your insights have led to many significant

improvements to our paper. We are honored to receive your positive endorsement for publication.