UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Propagation of societal gender inequality by internet search algorithms

Vlasceanu, M.; Amodio, D.M.

DOI
10.1073/pnas.2204529119

Publication date
2022

Document Version
Final published version

Published in
Proceedings of the National Academy of Sciences of the United States of America

License
CC BY

Link to publication

Citation for published version (APA):
Vlasceanu, M., & Amodio, D. M. (2022). Propagation of societal gender inequality by internet search algorithms. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(29), Article e2204529119. https://doi.org/10.1073/pnas.2204529119

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:12 Oct 2026

PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES OPEN ACCESS

Propagation of societal gender inequality by internet search algorithms

Madalina Vlasceanu[a,1] and David M. Amodio[a,b]

Edited by Susan Fiske, Princeton University, Princeton, NJ; received March 14, 2022; accepted May 27, 2022

Humans increasingly rely on artificial intelligence (AI) for efficient and objective decision-making, yet there is increasing concern that algorithms used by modern AI systems produce discriminatory outputs, presumably because they are trained on data in which societal biases are embedded. As a consequence, their use by human decision makers may result in the propagation, rather than reduction, of existing disparities. To assess this hypothesis empirically, we tested the relation between societal gender inequality and algorithmic search output and then examined the effect of this output on human decision-making. First, in two multinational samples ($n = 37$, 52 countries), we found that greater nation-level gender inequality was associated with more male-dominated Google image search results for the gender-neutral keyword "person" (in a nation's dominant language), revealing a link between societal-level disparities and algorithmic output. Next, in a series of experiments with human participants ($n = 395$), we demonstrated that the gender disparity associated with high- vs. low-inequality algorithmic outputs guided the formation of gender-biased prototypes and influenced hiring decisions in novel scenarios. These findings support the hypothesis that societal-level gender inequality is recapitulated in internet search algorithms, which in turn can influence human decision makers to act in ways that reinforce these disparities.

inequality | algorithm | AI | bias | gender

Significance

People often rely on artificial intelligence (AI) algorithms to increase their decision-making efficiency and objectivity, yet systemic social biases have been detected in these algorithms' outputs. We demonstrate that gender bias in a widely used internet search algorithm reflects the degree of gender inequality existing within a society. We then find that exposure to the gender bias patterns in algorithmic outputs leads people to think and act in ways that reinforce societal inequality, suggesting a cycle of bias propagation between society, AI, and users. These findings call for an integrative model of ethical AI that includes human psychological processes to illuminate the formation, operation, and mitigation of algorithmic bias.

Artificial intelligence (AI) algorithms have become pervasive in human decision-making. In virtually all sectors of society, from national security (1) and criminal justice (2) to medicine (3) and education (4), AI algorithms are used to improve the efficiency and accuracy of decisions. One preeminent application is to internet search engines, such as Google Search, which are used daily by people of all ages to gather information that informs their beliefs and behaviors (5). The transfer of such decision power from bias-prone humans to ostensibly objective algorithms has been celebrated as a step toward more fair and rational decision-making (6) and more intelligent problem solving (7).

Yet, despite the promise of objectivity by AI-based decision-making, there is mounting concern about systemic social biases in algorithmic output (8–12). Patterns of discrimination toward women and racial minorities have been detected in algorithms used for hiring decisions (13), delivery of job and housing advertisements (14, 15), university admissions decisions (4), criminal sentencing (2, 16), and health-care allocation (3), among others. For example, Facebook's job advertisement algorithm targeting users based on their gender was found to disproportionately suggest stereotypically feminine jobs (e.g., nurse or secretary) to women and stereotypically masculine jobs (e.g., janitor or taxi driver) to men, thus further propagating existing gender disparities in the labor market (14).

What causes such biases? A common assumption is that they reflect existing disparities in society represented in the data on which algorithms are trained (17), as well as other aspects of their development, from data collection to data preparation, model development, model evaluation, model postprocessing, and model deployment (18). To the extent algorithms, such as those used widely in internet search engines, encode preexisting social biases (19–21), there is a risk that their outputs could propagate these existing biases (22), thereby reinforcing and strengthening societal and economic inequality. This perspective suggests that AI decision tools could potentially contribute to a cycle of perpetuating inequality, whereby existing social biases are recapitulated in algorithmic outputs, which in turn guide human decisions that reinforce the existing inequality.

To begin to investigate this proposed cycle of bias propagation empirically, we asked whether the degree of inequality within a society relates to patterns of bias in algorithmic output and, if so, whether exposure to such output could influence human decision makers to act in accordance with these biases, thereby reinforcing them. To this end, we first assessed the relationship between existing societal inequality and algorithmic internet search output, in the context of gender, in two multinational studies. Then, we experimentally manipulated exposure to search outputs matching those

Author affiliations: [a]Department of Psychology, New York University, New York, NY 10003; and [b]Department of Psychology, University of Amsterdam, 1001 NK Amsterdam, The Netherlands

Author contributions: M.V. and D.M.A. designed research; M.V. and D.M.A. performed research; M.V. contributed new reagents/analytic tools; M.V. analyzed data; and M.V. and D.M.A. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Copyright © 2022 the Author(s). Published by PNAS. This open access article is distributed under Creative Commons Attribution License 4.0 (CC BY).

[1]To whom correspondence may be addressed. Email: vlasceanu@nyu.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2204529119/-/DCSupplemental.

Published July 12, 2022.

Downloaded from https://www.pnas.org by UNIVERSITEIT VAN AMSTERDAM (UVA) 215 on February 2, 2023 from IP address 146.50.146.158.

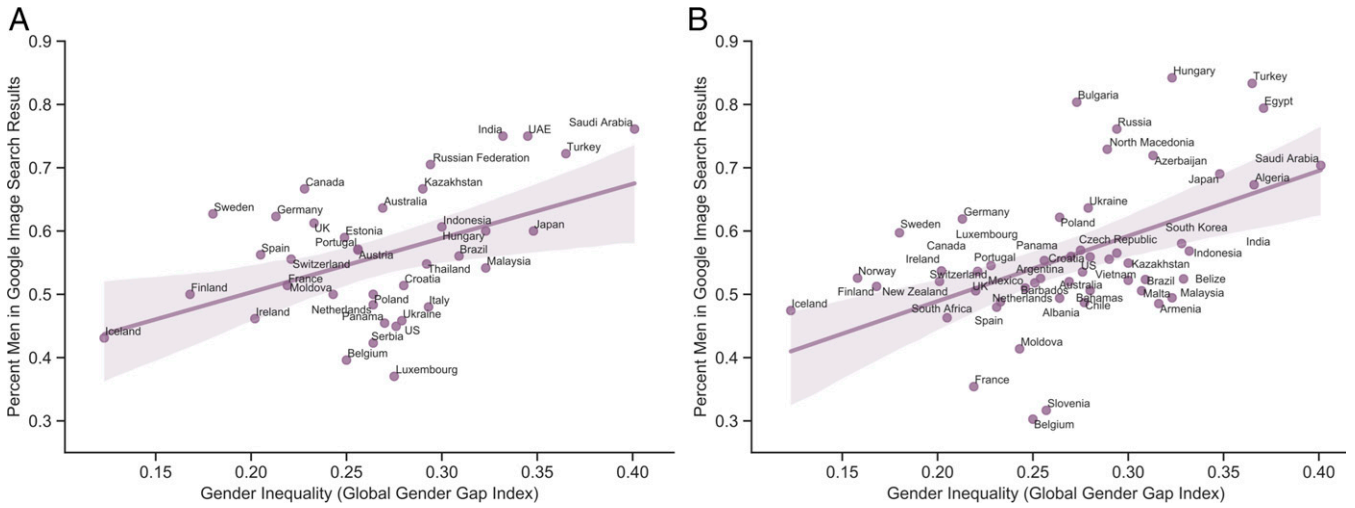

Fig. 1. Study 1 (*A*) and Study 2 (*B*) results. Gender inequality (derived from the GGGI computed by the World Economic Forum in 2020; *x* axis) as a function of gender bias in Google image search results, displayed here as the percentage of men in the results of Google images when searching for the gender-neutral keyword "person" in the language and with the IP address of each nation (*y* axis).

Downloaded from https://www.pnas.org by UNIVERSITEIT VAN AMSTERDAM (UVA) 215 on February 2, 2023 from IP address 146.50.146.158.

observed in high- and low-gender inequality nations to test whether they produce corresponding gender bias in human participants' cognitive concepts and decisions. We hypothesized that the algorithmic outputs of search engines track preexisting societal-level gender biases (H1) and that exposure to these biased algorithmic outputs shape users' cognitive concepts (H2) and decisions (H3) in manners consistent with societal gender inequality.

In Study 1, we tested the hypothesis that the degree of gender inequality within a society would predict a pattern of gender disparity in internet search output. To do so, we first obtained an open-source dataset known as the Global Gender Gap Index [GGGI (23)], containing rankings of gender inequality for 153 countries. The GGGI represents the magnitude of gender inequality in economic participation and opportunity, educational attainment, health and survival, and political empowerment for each of 153 nations, thus providing societal-level gender inequality scores for each country. Next, to develop an assessment of gender bias in algorithmic output, we drew upon classic research on the prototype bias in cognitive psychology (24, 25), whereby words that should refer with equal probability to a man or a woman, such as "person" (26), or "human" (27), are more often assumed to be a man (28). In recent research using word embeddings trained on massive internet text corpora, words representing the concept of "people" (e.g., "somebody" or "human") were more likely to cooccur with terms for "men" than for "women"—a demonstration of the male-default bias collectively displayed across individuals in a society (29).

This pervasive gender bias, documented for decades in human cognition, has also been observed in algorithmic outputs. For instance, online translation algorithms make gender-biased errors, assigning a masculine gender to professions such as "doctor" despite the original sentence implying a woman (30). Similarly, image search results for common professions such as "engineer" or "author" underrepresent the true proportion of women in those occupations (31, 32).

On the basis of such biases found in image search results, we developed an approach which involved conducting Google image searches for the keyword "person" within a nation (in its dominant local language) via VPN (virtual private networks; see *Methods* for additional details on this procedure). We chose the term "person" because of its single form even in gendered languages such as Spanish, in which the word for person is "*la persona*" regardless of the gender of the person. In contrast, gendered languages use distinct forms when referring to professionals; in Spanish, for instance, a male doctor is "*el doctor*," whereas a female doctor is "*la doctora*." Moreover, in addition to overcoming the constraint posed by gendered languages, the term "person" refers to an aspect of humanity—one's personhood—that is often undermined as a result of gender discrimination (33). We chose to conduct our search on Google given the overwhelming dominance of its search engine market share worldwide [i.e., over 92% (34)].

We collected data from this search in as many countries as possible. Given nation-specific restrictions and intermittent VPN access to local servers, at the time of Study 1 data collection, VPN connection was successful in 37 countries. A gender bias score was computed for each nation by coding the ratio of man to woman-presenting images within the first 100 results.

To test our central hypothesis, we assessed the relation between Global Gender Gap scores and the degree of gender disparity in image search algorithmic output. This effect was significant, $\beta = 0.472$, SE = 0.149, t (1,35) = 3.167, CI[0.169, 0.774], $P = 0.003$, such that the proportion of male images was higher in nations with greater gender inequality (computed as the inverse of the GGGI; Fig. 1*A*). This effect remained significant when adjusting for the percentage of women in the population of each nation, $\beta = 0.40$, SE = 0.16, t (2,34) = 2.47, CI[0.071, 0.729], $P = 0.019$ (*SI Appendix* for additional robustness checks). Thus, in a sample of 37 nations, we found that algorithmic gender bias tracks with societal gender inequality.

We replicated this procedure in Study 2, 3 mo after conducting Study 1, this time obtaining access to internet searches in 52 nations. Replicating Study 1 results, societal gender inequality related to the degree of gender disparity in image search algorithm output, $\beta = 0.50$, SE = 0.122, t (1,50) = 4.086, CI[0.25, 0.74], $P < 0.001$ (Fig. 1*B*), a pattern that remained significant when adjusting for the percentage of women in the population of each nation, $\beta = 0.49$, SE = 0.11, t (2,49) = 4.124, CI[0.25, 0.73], $P < 0.001$ (*SI Appendix* for additional robustness checks). Although the man/woman proportions in search output for nations represented in both studies ($n = 30$) differed slightly between the two data collection timepoints as search outputs naturally fluctuate over time (35), the proportions

at the two timepoints were highly correlated ($r = 0.60$, $P = 0.002$). Together, Studies 1 and 2 provide empirical evidence consistent with the proposal that societal-level gender disparities are reflected in algorithmic output.

Having demonstrated an association between societal gender inequality and algorithmic search output, we turned to our next question: whether exposure to such algorithmic outputs can shape people's cognitive concepts and decisions in manners consistent with preexisting societal inequality. Hence, in Study 3, we experimentally tested how exposure to gender ratios in image search results observed for low- versus high-inequality nations in Studies 1 and 2 affect human participants' judgment and decisions regarding novel categories.

To develop our approach, we built on seminal findings in the social categorization literature, which suggest that people naturally develop category prototypes as a way of assessing and predicting category membership (36). We focused on how such search results would influence judgments and decisions regarding novel categories, given preexisting associations between gender and the concept of personhood. Because novel categories are by definition devoid of such associations, they are optimal for experimentally isolating mechanisms of interest. Here, such mechanisms of interest are the formation of category prototypes—specifically, gender prototypes—and the impact of these prototypes on subsequent consequential decision-making. Thus, the goal of the following experimental investigation was to exemplify the effects of assigning men as the default members, or prototypes, of a category. We chose a labor-market scenario, given the contributing effects of economic participation and opportunity to global gender inequality (23) and in light of findings that algorithmic search results can shape occupational perceptions (31, 32).

In Study 3, 130 participants recruited in the United States (age: mean [M] = 42.1 y, SD = 15.9; 60% women, 39% men, <1% nonbinary) were presented with images depicting the Google image search results of four professions (chandler, draper, peruker, and lapidary). Because these professions are unfamiliar and thus novel to most Americans, our participants would have little prior knowledge about their gender distribution. The gender composition of each profession's image set was selected to represent the Google image search results for the keyword "person" for nations with the high gender inequality scores (approximately 90% men to 10% women in Hungary or Turkey) and those with low gender inequality scores (approximately 50% men to 50% women in Iceland or Finland) from Study 2 (Fig. 2). This resulted in conditions displaying a high ratio of men to women (9:1, high-inequality) or an equal ratio of men to women of (1:1, low-inequality) performing the profession of interest. Each profession type was randomly assigned to appear in the high-inequality or low-inequality condition, within subjects.

First, prior to viewing image search results, participants made baseline prototypicality judgments regarding each profession (e.g., "Who is more likely to be a peruker, a man or a woman?"). Replicating the classic gender prototype effect (37), participants in both conditions initially judged members of these professions as more likely to be a man (high-inequality: $\chi^2 = 21.06$, w = 0.05, $P < 0.001$; low-inequality: $\chi^2 = 12.94$, w = 0.04, $P < 0.001$; Fig. 3*A*), and this effect did not vary by participant gender (*SI Appendix*). Hence, prototypicality judgments did not differ by condition at baseline.

Next, following exposure to image search results, participants again rated the likelihood that a member of each profession was a man or woman. To test our central hypothesis that exposure to images representing the search results for "person" in high vs. low inequality nations influences gender prototypes, these gender prototypicality ratings were submitted to a generalized linear mixed effects model with condition (low vs. high inequality) and time (baseline vs. postmanipulation) as fixed effects and with by-participant and by-item (profession) random intercepts. Our hypothesis was supported by a significant condition by time interaction, $\beta = 0.23$, SE = 0.05, $t = 3.98$, Wald Z-test $P < 0.001$ (Fig. 3*A*): Exposure to search result patterns found in low-inequality nations reversed participants' male-biased prototypes relative to baseline, $\beta = 0.19$, SE = 0.04, $t = 4.76$, Wald Z-test $P < 0.001$, whereas exposure to search result patterns found in high-inequality nations appeared to reinforce and thus maintain this prototype, $\beta = 0.03$, SE = 0.03, $t = 0.97$, Wald Z-test $P = 0.32$ (Fig. 3*A*). Indeed, the postmanipulation prototype ratings in the low-inequality condition (mean = 0.58, SD = 0.49) were significantly more egalitarian (i.e., closer to the 0.5 benchmark) than those in the high-inequality condition (mean = 0.31, SD = 0.46) in a nonparametric statistical test (i.e., bootstrap, $P = 0.02$). The effect of condition on gender prototypicality judgment did not interact with participants' gender ($P = 0.29$).

Following prototype ratings, participants were asked to judge the likelihood that a man or woman would be hired in each profession ("What type of person is most likely to be hired as a

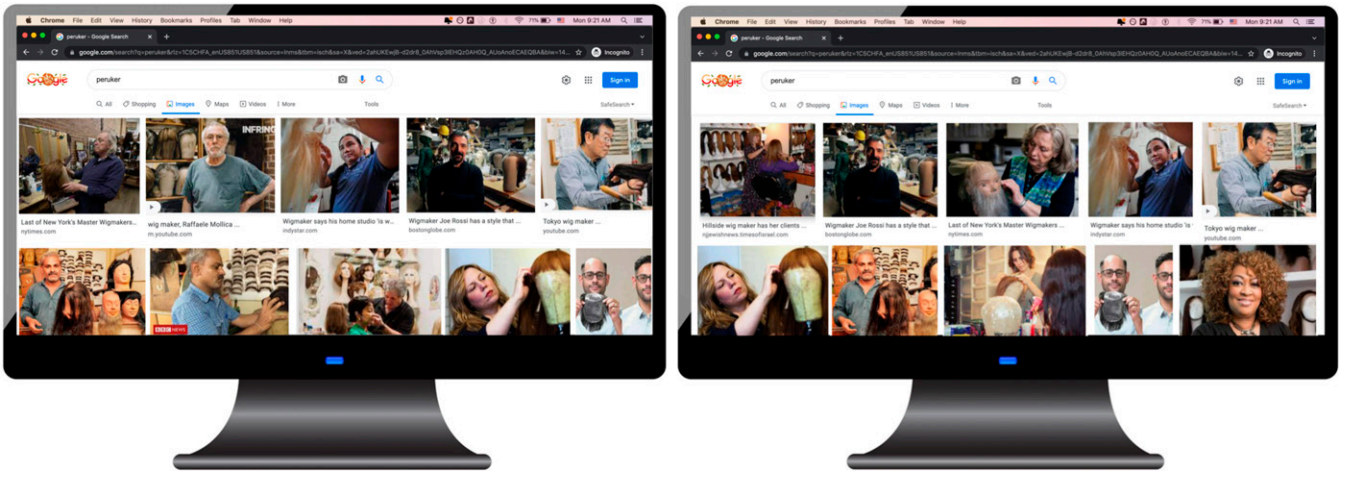

Fig. 2. Study 3 stimuli example. Google image search results for the search term "peruker" in the high-inequality condition (90% men, 10% women; *Left*) and the low-inequality condition (50% men, 50% women; *Right*).

Downloaded from https://www.pnas.org by UNIVERSITEIT VAN AMSTERDAM (UVA) 215 on February 2, 2023 from IP address 146.50.146.158.

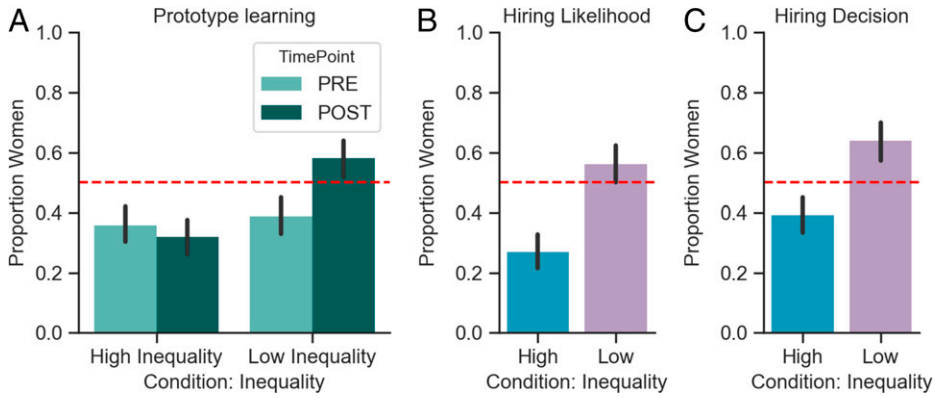

Fig. 3. Study 3 results. (*A*) Frequency of "Women" seen as prototypical of the novel category in the high- versus low-inequality conditions, at baseline (light green) and postmanipulation (dark green). (*B*) Proportion of women likely to be hired (postmanipulation) in the high-inequality (blue) condition (χ^2 = 55.3, w = 0.08, P < 0.001) and low-inequality (red) condition (χ^2 = 3.94, w = 0.02, P = 0.047). (*C*) Proportion of women hired (postmanipulation) in the high-inequality (blue) condition (χ^2= 12.06, w = 0.03, P < 0.001) and low-inequality (red) condition (χ^2 = 19.9, w = 0.05, P < 0.001). The red dash lines at y = 0.5 represent the chance level gender choice, and the χ^2 tests reported here compare the distributions to the 0.5 benchmark. Error bars represent 95% CI.

peruker?) and then, when presented with images of two job candidates (always one woman and one man) for a position in that profession, to make their own hiring choice ("Choose one of these applicants for a job as a peruker"). As expected, exposure to image sets in the low-inequality condition produced more egalitarian judgments of male vs. female hiring tendencies within a profession, β = 0.292, SE = 0.04, t = 7.25, Wald Z-test P < 0.001 (Fig. 3*B*), and a higher likelihood of choosing a woman job candidate, β = 0.24, SE = 0.041, t = 5.95, Wald Z-test P < 0.001 (Fig. 3*C*), compared with exposure to image sets in the high-inequality condition. In a mediation analysis (38), our overarching prediction—that exposure to high- versus low-inequality search results shape gender prototype formation, which in turn influences gender-biased hiring—was supported by a significant indirect effect (Fig. 4 and Tables 1 and 2).

To further probe the impact of search output exposure on prototype formation and decision-making, we next assessed the effects of biased algorithmic outputs in nations in which search results are biased in favor of women. Study 4 (n = 137, mean age = 41.81 y, SD = 15.17; 57.6% women) repeated the design of Study 3 but selected the composition of image sets for each profession to represent nations yielding the highest and lowest male:female ratios in the Study 2 Google image search results for the keyword "person" (approximately 90% men to 10% women in Hungary or Turkey, and approximately 30% men to 70% women in Belgium). Hence, the conditions in this study displayed a higher proportion of either men (9:1, majority-men condition) or women (7:3; majority-women condition) performing the profession of interest. We found that search output exposure biased the prototype formation symmetrically, suggesting a contrast effect: In the majority-men condition, the prototype became more male, β = 0.1058, SE = 0.038, t = 2.758, Wald Z-test P = 0.005, whereas in the majority-women condition, the prototype became more female, β = 0.3211, SE = 0.037, t = 8.466, Wald Z-test P < 0.001. This result conceptually replicates canonical prototype formation effects (38, 39) in the context of exposure to algorithmic outputs (*SI Appendix* for additional information).

Study 4 participants also more frequently selected men as more likely to be hired in a majority-men condition and women in a majority-women condition, β = 0.397, SE = 0.038, t = 10.40, Wald Z-test P < 0.001 (*SI Appendix* for additional information). An additional study (Study 5, n = 128; mean age = 39.6 y, SD age = 13.34; 61% women) replicated these results and further showed that exposure to majority-women image results increased decisions to hire women compared with exposure to majority-men search results (β = 0.33, SE = 0.041, t = 8.046, Wald Z-test P < 0.001; *SI Appendix* for additional information).

Discussion

Social biases have been widely observed in AI algorithms used for a variety of decision tools, and it has been proposed that such biases are rooted in existing societal disparities (8–12). Here, we provide empirical evidence in support of this proposal: In two cross-national studies of 37 (Study 1) and 52 (Study 2) countries, higher national levels of gender inequality predicted a larger male/female disparity in Google image search results for the gender-neutral search term "person." That is, societal-level gender inequality was reflected in search algorithm output.

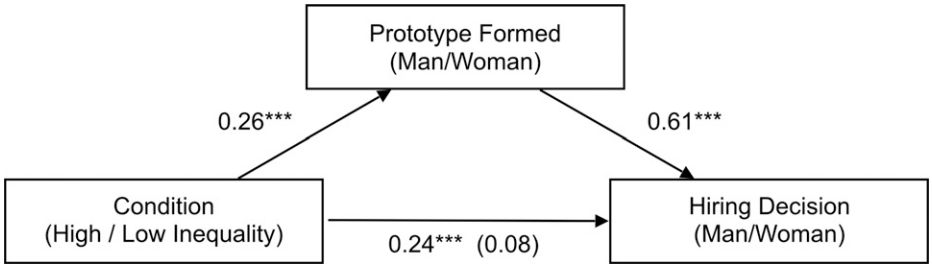

Fig. 4. Regression coefficients for the relationship between condition (high/low gender inequality in algorithmic search output) and hiring decision as mediated by the prototype formed. The standardized regression coefficient between condition and hiring decisions, controlling for the prototype formed, is in parentheses.

Downloaded from https://www.pnas.org by UNIVERSITEIT VAN AMSTERDAM (UVA) 215 on February 2, 2023 from IP address 146.50.146.158.

Table 1. Regression analyses for the mediation model

Predictors	*b* (SE)	*t*	*F*	df	R^2	*P*
Model 1						
Condition	0.24 (0.04)	5.78***	33.44	(1, 518)	0.058	<0.001
Model 2						
Condition	0.08 (0.03)	2.42*	181.8	(2, 517)	0.41	0.015
Prototype formed	0.61 (0.03)	17.6***			0.05	<0.001

b = regression coefficients; df = degrees of freedom. *$P < 0.05$; ***$P < 0.001$.

We then documented the effect of these search output disparities on human users' gender concepts and decisions in a novel labor-market scenario. When exposed to search result patterns from high-inequality nations in this context, participants were more likely to form male prototypes of a profession, expect men to be hired in that profession, and select men for a position in that profession. Exposure to search result patterns from low-inequality nations eliminated this effect: Participants were somewhat more likely to conceptualize, expect, and choose to hire more women. Mediation analyses indicated that gender disparities in search results influenced participant's gender prototypes, which in turn guided their hiring decisions. Finally, we found that exposure to search results depicting higher proportions of women fully reversed these preferences, with prototypes, expectations, and decisions strongly favoring women over men. Remarkably, a one-shot exposure to the biased search output was enough to generate correspondingly biased judgements and decisions. In everyday usage, however, people are likely exposed to many such algorithmic outputs, which may exert a cumulative impact on users' cognitive representations and their behavioral signatures.

Together, these findings demonstrate that societal levels of inequality are evident in internet search algorithms and that exposure to this algorithmic output can lead human users to think and potentially act in ways that reinforce this societal inequality. Our studies examined a specific test case of this pattern, in the context of gender inequality in image search algorithms and their experimental effects in novel decision domains; although real-world decisions are often more complex, this approach afforded a clear and valid proof of concept for the proposed mechanism of algorithmic bias propagation. However, algorithms vary widely in their functions, implementation, and usage. The extent to which they represent societal biases and influence users to act in accordance is also subject to variance. Nevertheless, evidence for algorithmic bias is already abundant (8–12), and its influence on human users, even when subtle, is likely to be pervasive (32).

Our findings further illuminate the impact of algorithmic bias on human cognition by demonstrating its effect on prototype formation. Given the widespread reliance on internet search engines in both formal education and daily life (40), people's cognitive concepts are likely shaped to a large extent by the outputs of computer algorithms (39). Our research experimentally demonstrates that when these outputs are biased, people's cognitive concepts (e.g., the prototypical person, the prototypical professional) can be shifted accordingly. As shown here and in prior work (41), these shifts can then trigger decisions congruent with a skewed reality that discriminate against those who do not fit one's conception of a social category's typical members. Moreover, when the mere concept of personhood overlaps mainly with one gender, the objectification of nonprototypical individuals can lead to additional discrimination (33) and oppression (42). Given the nation-level disparities observed in Studies 1 and 2, the effects of search algorithms on human users' concepts and worldviews may be disproportionately severe in societies with greater existing gender inequality.

It is notable that this research was conducted using a VPN-based internet search methodology. This approach allowed us to probe international differences in internet search algorithm function with low financial and administrative costs. However, national VPN access varies, such that it may be banned, obstructed, or otherwise unavailable in many locations (43) and thus may not provide an exhaustive international assessment. Nevertheless, this method provided us with access to 58 nations spanning six continents, with sufficient variability in national gender inequality to permit a valid test of our hypothesis. Thus, while VPN-based research may be limited to nonexhaustive cross-national assessments, it offers an effective tool for testing societal-level effects of internet-based algorithms.

Finally, by demonstrating a link between societal inequality, internet search algorithms, and human decision-making, our findings highlight potential intervention points for bias reduction efforts. Whereas past research has focused on interventions aimed at debiasing an algorithm's training set (44, 45) and elucidating the computations of deep neural network models (46), our findings suggest that to increase transparency and fairness in AI it will also be critical to uncover how human decision makers interact with and consume algorithmic output (47, 48). Psychology-based interventions aimed at breaking the cycle of bias propagation between society and AI might be particularly impactful. More broadly, our findings call for an integrative model of ethical AI that combines human psychology with computational and sociological approaches to illuminate the formation, operation, and mitigation of algorithmic bias.

Table 2. Causal mediation analyses: Nonparametric bootstrap CI, with 10,000 simulations

	Estimate	95% CI lower	95% CI upper	*P*
Indirect effect (ACME)	0.161	0.109	0.22	<0.001***
Direct effect (ADE)	0.084	0.015	0.16	0.018*
Total effect	0.246	0.163	0.33	<0.001***
Proportion mediated	0.656	0.470	0.92	<0.001***

ACME = average causal mediation effects; ADE = average direct effect. *$P < 0.05$; ***$P < 0.001$.

Downloaded from https://www.pnas.org by UNIVERSITEIT VAN AMSTERDAM (UVA) 215 on February 2, 2023 from IP address 146.50.146.158.

Methods

Methods Study 1.

Open science practices. Data can be found on the study's Open Science Framework (OSF) page. The data-analysis code (in Python) can be accessed as a Jupyter notebook on GitHub.

Measures. To compute a measure of algorithmic gender bias, a research assistant, blind to the study's hypotheses, coded the ratio of women- to men-presenting images of the first 100 results in Google image search using the gender-neutral keyword "person," translated (with Google Translate and using back-translation verification) in the language of each of the nations in which the search was conducted (Table 3). Of note, although some of these languages have gendered nouns (e.g., Spanish and Arabic) and others do not (e.g., English and Turkish), the gender proportion of search outputs obtained in gendered and nongendered languages did not differ in either Study 1 ($P = 0.093$) or Study 2 ($P = 0.597$). Moreover, the research assistant was instructed to conduct the search in all the 153 countries included in the 2020 Global Gender Gap Report (23). To conduct the search from different locations around the world, we used a VPN service (https://www.le-vpn.com) that activates Internet Protocol (IP) addresses corresponding to nations of interest. For each location activated, we used Google's IP address identifier to ensure the IP address under which each search was conducted was the intended one. In between each search, the research assistant cleared the browser history of the computers used to collect and code the data, in order to avoid contamination of the search results. Using this method, the VPN connection was successful in 37 nations (Table 3). The dataset was collected in August 2021.

The GGGI, published in the 2020 Global Gender Gap Report (23), was used as an index of societal-level gender inequality. The GGGI was introduced by the World Economic Forum in 2006 to capture the magnitude of gender disparity in economic participation and opportunity, educational attainment, health and survival, and political empowerment. According to these criteria, the index ranks each nation and provides them a score ranging from 0 to 1, 1 corresponding to full gender parity. For example, the highest score (0.82) was achieved by Iceland, meaning that this nation has closed 82% of its gender gap. We used this index to compute a measure of societal gender inequality, which we derived as one minus the GGGI index, such that higher values represent higher inequality.

Methods Study 2.

Open science practices. Data can be found on the study's OSF page. The data-analysis code (in Python) can be accessed as a Jupyter notebook on GitHub.

The same data collection procedure was followed as in Study 1. A different research assistant, also blind to the study's hypotheses, performed the image scraping and coding procedure. The dataset for Study 2 was collected in November 2021. The VPN connection was successful in 52 nations (Table 3). In total, across Studies 1 and 2, data were successfully collected from 58 nations (Fig. 5).

Methods Study 3.

Open science practices. The study's experimental design and hypotheses were preregistered at AsPredicted, and stimuli and data are available on the study's OSF page. The data-analysis code (in Python) can be accessed as a Jupyter notebook on GitHub.

Participants. We estimated that to obtain a power of 0.80 to detect a moderate effect size (Cohen's d) of 0.25 in a paired-sample comparison, a sample of 128 participants would be needed. Participants were recruited on Cloud Research (https://www.cloudresearch.com/), an Internet-based research platform similar to Amazon Mechanical Turk but with more intensive participant pool checks. Participants were compensated at the platform's standard rate (49). In total, we recruited 136 participants, of which 6 were excluded from the analysis on the basis of preregistered criteria (i.e., attention checks). The final sample contained 130 participants (age: M = 42.1 y, SD = 15.9; 60% women, 39% men, <1% nonbinary). The study protocol was approved by the New York University Institutional Review Board.

Stimuli materials. We designed "screenshots" of Google image search results for four professions (i.e., chandler, draper, peruker, and lapidary) of which participants were unlikely to have previous knowledge. For each profession, we designed screenshots of both men-dominated search results (high-inequality condition) and gender-balanced search results (low-inequality condition). In designing the stimuli,

Table 3. Language used to collect data from the countries represented in the dataset

Country	Search language	Search term
Albania[2]	Albanian*	Person
Algeria[2]	Arabic*	شخص
Argentina[2]	Spanish*	La persona
Armenia[2]	Armenian	մարդ
Australia[1,2]	English	Person
Azerbaijan[2]	Azerbaijani	şəxs
Austria[1]	German*	Person
Bahamas[2]	English	Person
Barbados[2]	English	Person
Belgium[1,2]	Dutch*	Persoon
Belize[2]	English	Person
Brazil[1,2]	Portuguese*	Pessoa
Bulgaria[2]	Bulgarian*	човек
Canada[1,2]	English	Person
Chile[2]	Spanish*	La persona
Croatia[1,2]	Croatian*	Osoba
Czech Republic[2]	Czech*	Osoba
Estonia[1]	Estonian	Isik
Egypt[1,2]	Arabic*	شخص
Finland[1,2]	Finnish	Henkilö
France[1,2]	French*	Personne
Germany[1,2]	German*	Person
Hungary[1,2]	Hungarian	Személy
Iceland[1,2]	Icelandic*	Manneskju
India[1,2]	Hindi*	व्यक्ति
Indonesia[1,2]	Indonesian	Orang
Ireland[1,2]	English	Person
Italy[1]	Italian*	Persona
Japan[1,2]	Japanese	人
Kazakhstan[1,2]	Kazakh	тұлға
Luxembourg[1,2]	Luxembourgish*	Persoun
Malaysia[1,2]	Malay*	Orang
Malta[2]	Maltese*	Persuna
Mexico[2]	Spanish*	La persona
Moldova[1,2]	Romanian*	Persoana
Netherlands[1,2]	Dutch*	Persoon
New Zealand[2]	English	Person
North Macedonia[2]	Macedonian*	лице
Norway[2]	Norwegian*	Person
Panama[1,2]	Spanish*	La persona
Poland[1,2]	Polish*	Osoba
Portugal[1,2]	Portuguese*	Pessoa
Russia[1,2]	Russian*	человек
Saudi Arabia[1,2]	Arabic*	شخص
Slovenia[2]	Slovenian*	Oseba
South Africa[2]	English	Person
Serbia[1]	Serbian*	особа
South Korea[2]	Korean	사
Spain[1,2]	Spanish*	La persona
Sweden[1,2]	Swedish*	Person
Switzerland[1,2]	German*	Person
Thailand[1]	Thai	บุคคล
Turkey[1,2]	Turkish	Kişi
Ukraine[1,2]	Ukrainian*	особа
United Arab Emirates[1]	Arabic*	شخص
United Kingdom[1,2]	English	Person
United States[1,2]	English	Person
Vietnam[2]	Vietnamese	người

Superscript numbers indicate the study in which each country appears. Asterisks indicate languages with gendered nouns.

Downloaded from https://www.pnas.org by UNIVERSITEIT VAN AMSTERDAM (UVA) 215 on February 2, 2023 from IP address 146.50.146.158.

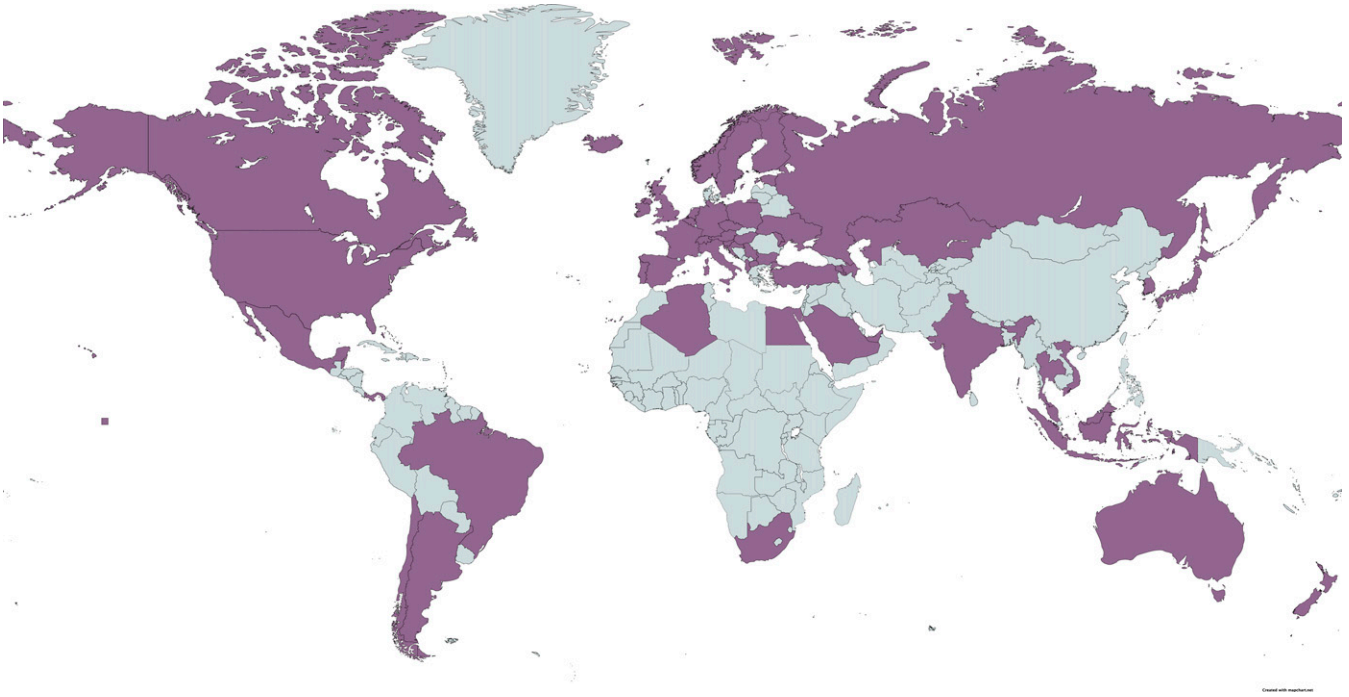

Fig. 5. Nations represented in the datasets of Studies 1 or 2 (i.e., in which the VPN connection was successful during at least one of the two data collection timepoints). Map created using https://www.mapchart.net/ (creative commons attribution license).

we used real images from Google image search, but we manipulated the proportion in which they appeared in the search result page.

Design and procedure. Data were collected in February 2022. Participants were told they would participate in a study about professions and were directed to the Qualtrics platform. After providing informed consent, participants were directed to the pretest phase in which they were asked a series of questions about each of the four professions (e.g., chandler), such as "What is a chandler?," "What is the age range of a person most likely to work as a chandler?," and "Who is more likely to be a chandler?" with options being "a man" or "a woman." Participants were also asked to estimate the salary range of a typical chandler, as well as the degree of friendliness and intelligence of the typical chandler. Then, participants were told they would see the screenshot of the Google image search results for the keyword "chandler," after which they were randomly assigned to either the high-inequality condition, in which the search results displayed 90% men and 10% women working as a chandler, or the low-inequality condition, in which the search results displayed 50% women and 50% men working as a chandler. These proportions were selected to represent the Study 2 Google image search results for the keyword "person" within nations with high gender inequality (approximately 90% men and 10% women in Hungary or Turkey), and low gender inequality (approximately 50% men and 50% women in Iceland or Finland). After seeing these image search results, participants were directed to the posttest phase, in which they were asked all the pretest questions again. Finally, participants were asked "What type of person is most likely to be hired as a chandler?," after which they entered a hiring decision task in which they were asked to choose between two applicants (a man and a woman) for each profession they learned about in the study. Participants were shown an image of each applicant and were asked to make a hiring decision in a forced-choice task. Finally, they completed a series of demographic questions and were debriefed.

Methods Study 4.

Open science practices. The study's experimental design and hypotheses were preregistered at AsPredicted. In addition, the stimuli and data can be found on the study's OSF page. The data-analysis code (in Python) can be accessed as a Jupyter notebook on GitHub.

Participants. We estimated that to obtain a power of 0.80 to detect a moderate effect size (Cohen's d) of 0.25 in a paired sample comparison, a sample of 128 participants would be needed. Participants were recruited on Cloud Research and were compensated at the platform's standard rate (49). In total, we recruited 139 participants, of which 2 were excluded from the analysis on the basis of preregistered criteria (i.e., attention checks). The final sample contained 137 participants (age: M = 41.81 y, SD = 15.17; 57.6% women, 42% men, <1% nonbinary). The study protocol was approved by the New York University Institutional Review Board.

Stimuli materials. We used the same materials as in Study 3, with one exception: The gender composition of images sets for each profession were now selected to represent the Study 2 Google image search results for the keyword "person" within nations with the highest search output proportion in favor of men (approximately 90% men and 10% women in Hungary or Turkey) and nations with the highest search output proportion in favor of women (approximately 30% men and 70% women in Belgium). This resulted in conditions displaying a higher proportion of either men (9:1) or women (7:3) performing the profession of interest. Each profession was randomly assigned to appear in the majority men or majority women condition, within subjects.

Design and procedure. The data were collected in December 2021. The design and procedure were the same as in Study 3, with one exception: In Study 4 we did not include the hiring decision task.

Methods Study 5.

Open science practices. The study's experimental design and hypotheses were preregistered at AsPredicted. In addition, the stimuli and data can be found on the study's OSF page. The data-analysis code (in Python) can be accessed as a Jupyter notebook on GitHub.

Participants. We estimated that to obtain a power of 0.80 to detect a moderate effect size (Cohen's d) of 0.25 in a paired sample comparison, a sample of 128 participants would be needed. Participants were recruited on Cloud Research and were compensated at the platform's standard rate (49). In total, we recruited 131 participants, of which 3 were excluded from the analysis on the basis of preregistered criteria (i.e., attention checks). The final sample contained 128 participants (age: M = 39.6 y, SD = 13.34; 61% women, 38% men, <1% nonbinary). The study protocol was approved by the New York University Institutional Review Board.

Stimuli materials. We used the same materials as in Study 4.

Design and procedure. The data were collected in January 2022. The design and procedure were the same as in Study 4, with one exception: In Study 5 we added a hiring decision task at the end of the study, in which participants were asked to choose between two applicants (a man and a woman) for the profession

Downloaded from https://www.pnas.org by UNIVERSITEIT VAN AMSTERDAM (UVA) 215 on February 2, 2023 from IP address 146.50.146.158.

in question. Participants were shown an image of each applicant and were asked to make a hiring decision in a forced choice task.

Data Availability. Study data, materials, and code have been deposited in OSF (https://osf.io/scvjk/) (50) and GitHub (https://github.com/mvlasceanu/Albias) (51).

ACKNOWLEDGMENTS. This work has been approved by the New York University (NYU) Institutional Review Board, and informed consent was obtained from all participants. This work was funded by the NYU Alliance for Public Interest Technology and the Netherlands Organisation for Scientific Research (VICI 016.185.058).

1. K. W. Bowyer, Face recognition technology: Security versus privacy. *IEEE Technol. Soc. Mag.* **23**, 9–19 (2004).
2. J. Larson, S. Mattu, L. Kirchner, J. Angwin, How we analyzed the COMPAS recidivism algorithm. *ProPublica* (2016). https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Accessed 23 June 2021.
3. Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
4. M. V. Santelices, M. Wilson, Unfair treatment? The case of Freedle, the SAT, and the standardization approach to differential item functioning. *Harv. Educ. Rev.* **80**, 106–134 (2010).
5. R. Epstein, R. E. Robertson, The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E4512–E4521 (2015).
6. M. K. Lee, Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data Soc.* **5**, 2053951718756684 (2018).
7. S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach* (Pearson, 2002)
8. M. Broussard, *Artificial Unintelligence: How Computers Misunderstand the World* (MIT Press, 2018).
9. K. Crawford, *The Atlas of AI* (Yale University Press, 2021).
10. A. Kadiresan, Y. Baweja, O. Ogbanufe, "Bias in AI-based decision-making" in *Bridging Human Intelligence and Artificial Intelligence*, M. V. Albert, L. Lin, M. J. Spector, L. S. Dunn, Eds. (Springer, Cham, Switzerland, 2022), pp. 275–285.
11. S. U. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York University Press, 2018).
12. C. O'neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown, 2016).
13. J. Dastin, Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (2018). https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. Accessed 11 February 2022.
14. M. Ali *et al.*, Discrimination through optimization: How Facebook's ad delivery can lead to biased outcomes. *Proc. ACM Hum. Comput. Interact.* **3**, 1–30 (2019).
15. K. Benner, G. Thrush, M. Isaac, Facebook engages in housing discrimination with its ad practices, U.S. says. *NY Times*, 28 March 2019.
16. J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias. *ProPublica* (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed 11 February 2022.
17. R. S. Baker, A. Hawn, Algorithmic bias in education. *Int. J. Artif. Intell. Educ.*, 10.1007/s40593-021-00285-9 (2021).
18. H. Suresh, J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle" in *Proceedings of the 2021 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Association for Computing Machinery, 2021), pp. 1–9.
19. A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
20. B. Friedman, H. Nissenbaum, Bias in computer systems. *ACM Trans. Inf. Syst.* **14**, 330–347 (1996).
21. N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E3635–E3644 (2018).
22. L. Paquette, J. Ocumpaugh, Z. Li, A. Andres, R. Baker, Who's learning? Using demographics in EDM research. *J. Educ. Data Min.* **12**, 1–30 (2020).
23. World Economic Forum, "Global Gender Gap Report 2020" (World Economic Forum, 2020).
24. N. M. Henley, "Molehill or mountain: What we know and don't know about the masculine as generic in English" in *Gender and Thought: Psychological Perspectives*, M. Crawford, M. Gentry, Eds. (Springer, 1989), pp. 59–78.
25. J. Silveira, Generic masculine words and thinking. *Womens Stud. Int. Q.* **3**, 165–178 (1980).
26. M. C. Hamilton, Masculine bias in the attribution of personhood: People = male, male = people. *Psychol. Women Q.* **15**, 393–402 (1991).
27. A. H. Bailey, M. LaFrance, Anonymously male: Social media avatar icons are implicitly male and resistant to change. *Cyberpsychology* **10**, 10.5817/CP2016-4-8 (2016).
28. A. H. Bailey, M. LaFrance, J. F. Dovidio, Is man the measure of all things? A social cognitive account of androcentrism. *Pers. Soc. Psychol. Rev.* **23**, 307–331 (2019).
29. A. H. Bailey, A. Williams, A. Cimpian, Based on billions of words on the internet, people = men. *Sci. Adv.* **8**, eabm2463 (2022).
30. G. Stanovsky, N. A. Smith, L. Zettlemoyer, Evaluating gender bias in machine translation. arXiv [Preprint] (2019). https://arxiv.org/abs/1906.00591 (Accessed 12 February 2022).
31. M. Kay, C. Matuszek, S. A. Munson, "Unequal representation and gender stereotypes in image search results for occupations" in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2015), pp. 3819–3828.
32. D. Metaxa, M. A. Gan, S. Goh, J. Hancock, J. A. Landay, An image of society: Gender and racial representation and impact in image search results for occupations. *Proc. ACM Hum. Comput. Interact.* **5**, 1–23 (2021).
33. J. P. Leyens *et al.*, The emotional side of prejudice: The attribution of secondary emotions to ingroups and outgroups. *Pers. Soc. Psychol. Rev.* **4**, 186–197 (2000).
34. StatCounter, Search engine market share worldwide (2022). https://gs.statcounter.com/search-engine-market-share. Accessed 2 March 2022.
35. J. E. Pitkow *et al.*, Personalized search. *Commun. ACM* **45**, 50–55 (2002).
36. S. T. Fiske, S. E. Taylor, *Social Cognition* (McGraw-Hill Book Company, 1991).
37. E. H. Rosch, Natural categories. *Cognit. Psychol.* **4**, 328–350 (1973).
38. D. Tingley, T. Yamamoto, K. Hirose, L. Keele, K. Imai, Mediation: R package for causal mediation analysis (2014).
39. A. Lenhart, M. Simon, M. Graziano, "The Internet and Education: Findings of the Pew Internet & American Life Project" (Pew Internet and American Life Project, Washington, DC, 2001).
40. N. Cantor, W. Mischel, "Prototypes in person perception" in *Advances in Experimental Social Psychology*, L. Berkowitz, Ed. (Academic Press, 1979), vol. **12**, pp. 3–52.
41. M. A. Hogg, A social identity theory of leadership. *Pers. Soc. Psychol. Rev.* **5**, 184–200 (2001).
42. S. J. Gervais, P. Bernard, O. Klein, J. Allen, "Toward a unified theory of objectification and dehumanization" in *Objectification and (De)Humanization*, S. J. Gervais, Ed. (Springer, 2013), pp. 1–23.
43. Y. Liu, C. A. Shue, "Avoiding VPN bottlenecks: Exploring network-level client identity validation options" in *International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness* (Springer, Cham, Switzerland, 2021), pp. 280–300.
44. T. Baer, V. Kamalnath, *Controlling Machine-Learning Algorithms and Their Biases* (McKinsey Insights, 2017).
45. T. Gebru *et al.*, "Excerpt from datasheets for datasets" in *Ethics of Data and Analytics*, K. Martin, Ed. (Auerbach Publications, 2022), pp. 148–156.
46. M. Du, F. Yang, N. Zou, X. Hu, Fairness in deep learning: A computational perspective. *IEEE Intell. Syst.* **36**, 25–34 (2020).
47. L. P. Gleaves, R. Schwartz, D. A. Broniatowski, The role of individual user differences in interpretable and explainable machine learning systems. arXiv [Preprint] (2020). https://arxiv.org/abs/2009.06675 (Accessed 3 March 2022).
48. M. Nourani *et al.*, "Anchoring bias affects mental model formation and user reliance in explainable AI systems" in *26th International Conference on Intelligent User Interfaces* (Association for Computing Machinery, 2021), pp. 340–350.
49. J. Chandler, C. Rosenzweig, A. J. Moss, J. Robinson, L. Litman, Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behav. Res. Methods* **51**, 2022–2038 (2019).
50. M. Vlasceanu, D. M. Amodio, Propagation of societal gender inequality by internet search algorithms. Open Science Framework. https://osf.io/scvjk/. Deposited 11 March 2022.
51. M. Vlasceanu, D. M. Amodio, Data for "Propagation of societal gender inequality by internet search algorithms." GitHub. https://github.com/mvlasceanu/Albias. Deposited 11 March 2022.

Downloaded from https://www.pnas.org by UNIVERSITEIT VAN AMSTERDAM (UVA) 215 on February 2, 2023 from IP address 146.50.146.158.