Algorithms propagate gender bias in the marketplace—with consumers' cooperation

Web Appendix

Table of Contents

Appendix S1: Procedures for Detecting Marketplace Gender Bias in Text

Procedures. Once we have converted a word into a vector, we can mathematically use it in our analysis. The simplest thing we can do is to find the similarity between two words. Just as in any unsupervised learning method, the distance between words is used to measure similarity. However, unlike in 2 or 3 dimensions when we use Euclidean distance, in this high-dimensional space, we use the cosine of the angle between word vectors to measure similarity.

$$Cosine\ Similarity = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2}\sqrt{\sum_{i=1}^{n} {y_i}^2}}$$

Values of $cos\theta$ can vary from +1 to -1 because cos (0) = 1, cos(90) = 0, and cos(180) = -1. For words that are very similar, cosine distance will be near 1. As words become more dissimilar, cosine distance will decrease and become nearly 0 (moving from 1 to 0), whereas when words are opposite in meaning, it will become negative (moving from 0 to -1).

The method to detect gender bias in the text can be described in the following manner. Let *X* represent a set of male words and *Y* a set of female words. Let *A* be the set containing positive attributes and *B* the set for negative attributes.

1. Using cosine similarity, we find the similarity of each target female-related and male-related word with each positive and negative attribute word. Therefore, for a male word *x* (member of set X) and an attribute word *a* (member of set *A* containing positive attributes) cos (*x, a*) will give us the value of the similarity between *x* and *a,* where $cos\ cos\ (x, a) = \frac{x.a}{\|x\|\|a\|}$.
2. We then calculate the average similarity between the target word *x* and all the positive attributes in set *A* as $mean_{a\epsilon A} cos\ (x, a)$. The net similarity to positive attributes is $mean_{a\epsilon A} cos\ (x, a)$ - $mean_{b\epsilon B} cos\ (x, b)$. If $mean_{a\epsilon A} cos\ (x, a)$ - $mean_{b\epsilon B} cos\ (x, b) > 0$, it shows that x is closer to positive rather than negative attributes. If $S(x, A, B) = mean_{a\epsilon A} cos\ (x, a)$ - $mean_{b\epsilon B} cos\ (x, b)$, then for set *X*, $\sum_{x\in X} S(x, A, B)$ captures the sum of this net similarity to positive attributes for all of its members. In the same vein, the net similarity of all the members of set *Y* (i.e., set of female words) to positive and negative attributes is given by $\sum_{y\in Y} S(y, A, B)$. $\sum_{x\in X}\ mean_{a\epsilon A} cos\ (x, a)$ is the semantic similarity of male words with positive attributes and $\sum_{x\in X}\ mean_{a\epsilon A} cos\ (x, b)$ is the semantic similarity of male words with negative attributes.
3. The main measure of a marketplace gender bias would then be $S(X, Y, A, B) = \sum_{x\in X}\ S(x, A, B) - \sum_{y\in Y}\ S(y, A, B)$. A positive value of $S(X, Y, A, B)$ would show that names in set *X* (male words) are more similar to positive attributes than names in set *Y* (female words). However, a negative value would show that words in set *Y* are closer to positive attributes than those in set *X.*
4. However, $S(X, Y, A, B)$ is just one measure of relative similarity, and one could argue that there is no statistically significant difference between $\sum_{x\in X} S(x, A, B)$ and $\sum_{y\in Y} S(y, A, B)$. Therefore, we need to rule out the null hypothesis that $\sum_{x\in X} S(x, A, B) = \sum_{y\in Y} S(y, A, B)$. Akin to any two-tailed hypothesis testing, ruling out this null hypothesis requires estimating the probability of not being able to reject the null hypothesis (which is

captured via *p* value).

Such probability can be calculated by 1) obtaining the similarity score for various partitions of the given names in two sets by creating sets like *X* = {he, her, boy} and *Y*= {she, his, girl}, and 2) finding the number of times such partitions give a score more extreme than the obtained score $S\ (X, Y, A, B)$. This is a non-parametric permutation test.

Formally, if (X_i, Y_i) represents the potential random shuffling of words in set *X* and *Y*, then the probability of not being able to reject the null hypothesis will be

$$Probability = \frac{Number\ of\ times\ (S(X_i, Y_i, A, B) > |S(X, Y, A, B)|)}{Number\ of\ all\ the\ potential\ shuffling\ of\ set\ X\ and\ Y}$$

This will give us a two-tailed *p*-value. It is important to note that the denominator of the probability estimate can get very large if we have many members in set *X* and *Y*. In such situations, a sufficiently high number of shuffles provide an approximate value of probability. For each comparison reported in this work, we did 5000 shuffles of names in the two sets.

5. Finally, the effect size of the marketplace gender bias can be estimated by

$$\frac{mean_{x \in X} S(x,A,B) - mean_{y \in Y} S(y,A,B)}{StandardDeviation_{w \in X \cup Y} S(w,A,B)}$$

This is a normalized estimate of marketplace gender bias similar to Cohen's *d*.

Appendix S2: Psychographic Attributes Literature

Selected Research Using Psychographic Traits for Consumers' Personality

Type of Trait	Relation to Big Five Model	Psychographic Traits	Journal	Field/Area	Authors(s) and Year
Positive	Openness to Experience	Innovative (+), Loyal (-), Creative (+)	Journal of Marketing	Marketing	Ailawadi, Neslin, & Gedenk (2001)
			Journal of Marketing Research		Steenkamp & Maydeu-Olivares (2015)
			International Journal of Management and Business Research		Vazifehdoost, Akbari, & Charsted (2012)
	Conscientiousness	Rational (+), Logical (+), Planned (+), Thorough (+), Disciplined (+), Dependable (+), Reliable (+)	Journal of Marketing	Marketing, Consumer Psychology	Ailawadi, Neslin, & Gedenk (2001)
			Journal of Marketing Research		Steenkamp & Maydeu-Olivares (2015)
			International Journal of Management and Business Research		Vazifehdoost, Akbari, & Charsted (2012)
			Journal of Applied Psychology		Mount, Barrick, and Strauss (1994)
	Extraversion	Jolly (+), Industrious (+)	Psychology & Marketing	Marketing, Consumer Psychology	Clark & Goldsmith (2005)
			Journal of Marketing Research		Steenkamp & Maydeu-Olivares (2015)
			International Journal of Management and Business Research		Vazifehdoost, Akbari, & Charsted (2012)
	Agreeableness	Kind (+)	Journal of Marketing Research	Marketing, Consumer Psychology	Steenkamp & Maydeu-Olivares (2015)
			International Journal of Management and Business Research		Vazifehdoost, Akbari, & Charsted (2012)

Type of Trait	Relation to Big Five Model	Psychographic Traits	Journal	Field/Area	Authors(s) and Year
	Neuroticism	Certain (-), Resisted (-), Relaxed (-)	Journal of Marketing	Marketing, Consumer Psychology	Ailawadi, Neslin, & Gedenk (2001)
			Psychology & Marketing		Dholakia (2000)
			Journal of Marketing Research		Steenkamp & Maydeu-Olivares (2015)
			International Journal of Management and Business Research		Vazifehdoost, Akbari, & Charsted (2012)
Negative	Openness to Experience	Risky (+), Rigid (-), Intolerant (-)	Journal of Consumer Research	Marketing, Consumer Psychology	Raju (1980)
			Journal of Marketing Research		Steenkamp & Maydeu-Olivares (2015)
			Journal of Consumer Psychology		Wood & Neal (2009)
	Conscientiousness	Frivolous (-), Emotional (-), Unreliable (-), Irresponsible (-), Indulgent (-)	Journal of Marketing	Marketing, Consumer Psychology	Ailawadi, Neslin, & Gedenk (2001)
			Psychology & Marketing		Evers, Gruner, Sneddon, and Lee (2018)
			Journal of Applied Psychology		Mount, Barrick, and Strauss (1994)
			International Journal of Management and Business Research		Vazifehdoost, Akbari, & Charsted (2012)
			Sociological Theory		Wherry (2008)
	Extraversion	Submissive (-)	Journal of Consumer Psychology	Consumer Psychology	Moon (2002)
	Agreeableness	Conformist (+), Vain (+), Unkind (-), Sarcastic (-), Unfriendly (-)	Journal of Marketing	Marketing, Consumer Psychology,	Ailawadi, Neslin, & Gedenk (2001)
			Journal of Marketing Research		Steenkamp & Maydeu-Olivares (2015)
			Journal of Consumer Research		Netemeyer, Burton, & Lichtenstein (1995)

Type of Trait	Relation to Big Five Model	Psychographic Traits	Journal	Field/Area	Authors(s) and Year
			International Journal of Management and Business Research		Vazifehdoost, Akbari, & Charsted (2012)
			Journal of Consumer Research		Warren, Barsky, and McGraw (2018)
	Neuroticism	Impulsive (+), Indulgent (+), Tempted (+), Hedonistic (+), Fickle (+), Irritable (+), Emotional (+), Sensitive (+), Vindictive (+), Moody (+)	Journal of Marketing	Marketing, Consumer Psychology	Ailawadi, Neslin, & Gedenk (2001)
			Psychology & Marketing		Dholakia (2000)
			Journal of Retailing & Consumer Services		Kapoor, Balaji, Maity, & Jain (2021)
			Journal of Retailing & Consumer Services		Tarka, Kukar-Kinney, & Harnish (2022)
			Journal of Marketing Research		Steenkamp & Maydeu-Olivares (2015)
			International Journal of Management and Business Research		Vazifehdoost, Akbari, & Charsted (2012)
			Sociological Theory		Wherry (2008)

Marketing and Consumer Psychology Citations:

1. Steenkamp, J. B. E., & Maydeu-Olivares, A. (2015). "Stability and change in consumer traits: Evidence from a 12-year longitudinal study, 2002–2013." *Journal of Marketing Research*, 52(3), 287–308.
2. Ailawadi, K. L., Neslin, S.A., & Gedenk, K. (2001, January), "Pursuing the value-conscious consumer: Store brands versus national brand promotions." *Journal of Marketing*, 65, 71–89.
3. Raju, P.S. (1980), "Optimum stimulation level: Its relationship to personality, demographics, and exploratory behavior," *Journal of Consumer Research*, 7(3), 272–282.
4. Netemeyer, R. G., Burton, S., & Lichtenstein, D.R. (1995). "Trait aspects of vanity: Measurement and relevance to consumer behavior," *Journal of Consumer Research*, 21 (4), 612–626.
5. Wood, W., & Neal, D. T. (2009). "The habitual consumer." *Journal of Consumer Psychology,* 19(4), 579–592.
6. Evers, U., Gruner, R. L., Sneddon, J., & Lee, J. A. (2018). "Exploring materialism

and frugality in determining product end-use consumption behaviors." *Psychology & Marketing*, 35(12), 948–956.
7. Wherry, F. F. (2008). "The social characterizations of price: The fool, the faithful, the frivolous, and the frugal." *Sociological Theory*, 26(4), 363–379.
8. Vazifehdoost, H., Akbari, M., & Charsted, P. (2012). "The role of psychological traits in market mavensim using Big Five model." *International Journal of Management and Business Research*, 2(3), 243–252.
9. Mount, M. K., Barrick, M. R., & Strauss, J. P. (1994). "Validity of observer ratings of the Big Five personality factors." *Journal of Applied Psychology*, 79(2), 272.
10. Moon, Y. (2002). "Personalization and personality: Some effects of customizing message style based on consumer personality." *Journal of Consumer Psychology*, 12(4), 313–325.
11. Kapoor, P. S., Balaji, M. S., Maity, M., & Jain, N. K. (2021). "Why consumers exaggerate in online reviews? Moral disengagement and dark personality traits." *Journal of Retailing and Consumer Services*, 60, 102496.
12. Warren, C., Barsky, A., & McGraw, A. P. (2018). "Humor, comedy, and consumer behavior." *Journal of Consumer Research*, 45(3), 529–552.
13. Tarka, P., Kukar-Kinney, M., & Harnish, R. J. (2022). "Consumers' personality and compulsive buying behavior: The role of hedonistic shopping experiences and gender in mediating-moderating relationships." *Journal of Retailing and Consumer Services*, 64, 102802.
14. Dholakia, U. M. (2000). "Temptation and resistance: An integrated model of consumption impulse formation and enactment." *Psychology & Marketing*, 17(11), 955–982.

Original Citations *(traits/words used to create the dictionary were taken from):*

1. Aaker, J. L. (1997). "Dimensions of brand personality." *Journal of Marketing Research*, 34(3), 347–356.
2. Anderson, N. H. (1968). "Likableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology*, 9(3), 272.
3. Berry, D. S., & McArthur, L. Z. (1985). "Some components and consequences of a babyface." *Journal of Personality and Social Psychology*, 48(2), 312.
4. Briggs, S. R. (1992). "Assessing the 5-factor model of personality description." *Journal of Personality*, 60(2), 253–293.
5. Eysenck, M. (2012). *Attention and arousal: Cognition and performance*. Springer Science & Business Media.

Bold Traits below are cited in the table (38 out of 59 traits have been directly cited in marketing contexts).

Positive trait word: Honest, reasonable, independent, **thorough**, **dependable**, **rational**, **relaxed**, **loyal**, **reliable**, **disciplined**, patience, **creative**, **innovative**, **planned**, resolute, **resisted**, **industrious**, **certain**, determined, wise, tough, **jolly**, civilized, strong, enterprising, quick, **logical**, original, methodical, **kind**

Negative trait words: **unfriendly**, **unkind**, **rigid**, **moody**, **intolerant**, **hedonistic**, **tempted**, fragile, **indulgent**, **irresponsible**, instinctive, dissatisfied, **conformist**, **impulsive**, **fickle**, **unreliable**, **emotional**, **vain**, lazy, **submissive**, **risky**, **irritable**, **frivolous**, inhibited, **sensitive**, **vindictive**, complicated, changeable, **sarcastic**

Gender pronoun and consumer psychographic attribute dictionaries. Based on prior literature (e.g., Garg et al. 2018), we applied the male/female target gender words shown below within the text analyses. We also drew on prior publications in marketing to develop a list of psychographic attributes, shown below, that are used for customer segmentation and targeting. Specifically, the psychographic attributes were drawn from: Anderson, 1968; Berry and McArthur, 1985; Eysenck, 1982; Hofstee and Raad, & Goldberg, 1992; Roberts, Wood, and Smith, 2005, Steenkamp and Maydeu-Olivares 2015. Then, extending previous research in computer science (e.g., Garg et al. 2018), we subsequently categorized these attributes into those that were desirable (positive) or undesirable (negative). A pretest presented in Appendix S3 validated this characterization of the attributes. Additionally, a sensitivity analysis presented in Appendix S6 ensured that the findings were robust to changes in the dictionary.

Female/Male target words:
Female target words: she, hers, her, woman, female, herself, women, females, gal, girl
Male target words: he, his, him, man, male, himself, men, males, guy, boy

Positive/Negative attribute words:
Positive attribute words: honest, reasonable, independent, thorough, dependable, rational, relaxed, loyal, reliable, disciplined, patience, creative, innovative, planned, resolute, resisted, industrious, certain, determined, wise, tough, jolly, civilized, strong, enterprising, quick, logical, original, methodical, kind
Negative attribute words: unfriendly, unkind, rigid, moody, intolerant, hedonistic, tempted, fragile, indulgent, irresponsible, instinctive, dissatisfied, conformist, impulsive, fickle, unreliable, emotional, vain, lazy, submissive, risky, irritable, frivolous, inhibited, sensitive, vindictive, complicated, changeable, sarcastic

Appendix S3: Attributes Pretest

Two hundred and seventy-two participants (age M = 34.88 years, 29% females) from Amazon Mechanical Turk (MTurk) completed a pretest to evaluate the attributes used in the manuscript. Each participant was randomly shown 10 attributes (out of 59). For each attribute, we asked the respondents to answer two questions (categories). Specifically, they were asked to categorize each attribute as negative–positive (options provided: negative, positive, or could be both). We also asked participants to categorize each attribute as undesirable–desirable (options provided: undesirable, desirable, or could be both). Each attribute received an average of 46 ratings on the two scales. As the first step for the negative–positive question, we calculated the proportion of responses that were categorized as positive for each attribute. The median split (0.55) was used to categorize each attribute into negative and positive categories. Attributes below the median were categorized as negative (ranged from 0.10 to 0.48), while attributes above the median were categorized as positive attributes (ranged from 0.55 to 0.93). Attributes (Median = 0.52) were similarly categorized as desirable (ranged from 0.52 to 0.95) versus undesirable (ranged from 0.17 to 0.49). The results show consistent categorization of attributes for the two categories. The results hold when a mean split rather than a median split is used.

In addition, we performed the analysis separately for male versus female participants (raters) to examine if the ratings differed depending on the gender of the rater. Similar to the attribute categories, we first calculated the proportion of responses that were categorized as positive by male participants (raters) only. The median split (0.57) was used to categorize each attribute into negative and positive categories. Second, we calculated the proportion of responses that were categorized as positive by female participants (raters) only. The median split (0.50) was used to categorize each attribute into negative and positive categories. Attributes below the median were categorized as negative, while attributes above the median were categorized as positive. The results showed no changes in the categorization of the dictionary into negative versus positive categories when only male versus female raters were used. Attributes for male (Median = 0.52) versus female (Median = 0.53) raters were similarly categorized into desirable versus undesirable categories. The results showed consistent categorization of the dictionary into undesirable versus desirable categories when only male versus female raters were used.

Based on the overall pretest responses, we show below the categorization of the dictionary into positive–negative and desirable–undesirable categories for each attribute in increasing order. Also, we show the categorization of the dictionary based on only male versus female raters.

Overall Responses

Attribute	Proportion of Positive Response	Categorization
dissatisfied	0.11	Negative
unkind	0.17	Negative
unfriendly	0.19	Negative
vain	0.25	Negative

irresponsible	0.26	Negative
intolerant	0.26	Negative
emotional	0.28	Negative
unreliable	0.28	Negative
irritable	0.30	Negative
rigid	0.30	Negative
fragile	0.30	Negative
vindictive	0.31	Negative
sarcastic	0.31	Negative
lazy	0.35	Negative
complicated	0.36	Negative
impulsive	0.37	Negative
fickle	0.39	Negative
frivolous	0.39	Negative
indulgent	0.40	Negative
risky	0.41	Negative
inhibited	0.43	Negative
hedonistic	0.44	Negative
submissive	0.45	Negative
tempted	0.46	Negative
changeable	0.46	Negative
instinctive	0.46	Negative
conformist	0.47	Negative
moody	0.48	Negative
sensitive	0.48	Negative
tough	0.55	Positive
resistant	0.57	Positive
certain	0.65	Positive
dependable	0.65	Positive
methodical	0.65	Positive
determined	0.66	Positive
quick	0.72	Positive
industrious	0.72	Positive
rational	0.76	Positive
wise	0.77	Positive
independent	0.79	Positive
creative	0.79	Positive
civilized	0.79	Positive
disciplined	0.80	Positive
honest	0.83	Positive
planned	0.84	Positive
resolute	0.85	Positive

logical	0.86	Positive
relaxed	0.86	Positive
jolly	0.87	Positive
enterprising	0.87	Positive
original	0.89	Positive
loyal	0.89	Positive
innovative	0.89	Positive
reasonable	0.90	Positive
reliable	0.90	Positive
thorough	0.90	Positive
kind	0.90	Positive
patient	0.91	Positive
strong	0.93	Positive

Responses by Male Raters

Attribute	**Proportion of Positive Response**	**Categorization**
dissatisfied	0.10	Negative
unkind	0.13	Negative
unfriendly	0.20	Negative
intolerant	0.27	Negative
emotional	0.28	Negative
vain	0.28	Negative
vindictive	0.30	Negative
fragile	0.31	Negative
irresponsible	0.31	Negative
rigid	0.34	Negative
unreliable	0.34	Negative
sarcastic	0.39	Negative
irritable	0.39	Negative
complicated	0.40	Negative
lazy	0.44	Negative
indulgent	0.45	Negative
risky	0.46	Negative
impulsive	0.46	Negative
frivolous	0.48	Negative
hedonistic	0.48	Negative
fickle	0.49	Negative
conformist	0.49	Negative
inhibited	0.50	Negative

submissive	0.51	Negative
instinctive	0.52	Negative
changeable	0.54	Negative
sensitive	0.55	Negative
tempted	0.55	Negative
moody	0.55	Negative
tough	0.57	Positive
determined	0.61	Positive
methodical	0.63	Positive
resistant	0.64	Positive
dependable	0.70	Positive
certain	0.71	Positive
industrious	0.72	Positive
creative	0.77	Positive
independent	0.78	Positive
quick	0.78	Positive
rational	0.80	Positive
jolly	0.81	Positive
wise	0.82	Positive
enterprising	0.83	Positive
civilized	0.84	Positive
resolute	0.84	Positive
innovative	0.84	Positive
loyal	0.85	Positive
relaxed	0.85	Positive
planned	0.86	Positive
disciplined	0.87	Positive
reliable	0.88	Positive
thorough	0.88	Positive
kind	0.88	Positive
reasonable	0.89	Positive
logical	0.89	Positive
honest	0.91	Positive
original	0.91	Positive
strong	0.93	Positive
patient	0.93	Positive

Responses by Female Raters

Attribute	Proportion of Positive Response	Categorization
dissatisfied	0.12	Negative
unfriendly	0.18	Negative
irritable	0.21	Negative
irresponsible	0.21	Negative
unreliable	0.21	Negative
unkind	0.21	Negative
vain	0.22	Negative
sarcastic	0.23	Negative
intolerant	0.25	Negative
lazy	0.25	Negative
rigid	0.26	Negative
impulsive	0.27	Negative
emotional	0.28	Negative
fragile	0.29	Negative
frivolous	0.29	Negative
fickle	0.29	Negative
vindictive	0.31	Negative
complicated	0.31	Negative
indulgent	0.35	Negative
inhibited	0.35	Negative
risky	0.36	Negative
changeable	0.37	Negative
tempted	0.37	Negative
submissive	0.39	Negative
hedonistic	0.40	Negative
instinctive	0.40	Negative
moody	0.40	Negative
sensitive	0.41	Negative
conformist	0.44	Negative
resistant	0.50	Negative
tough	0.53	Positive
certain	0.58	Positive
dependable	0.60	Positive
quick	0.65	Positive

methodical	0.67	Positive
determined	0.71	Positive
rational	0.71	Positive
disciplined	0.72	Positive
industrious	0.72	Positive
wise	0.72	Positive
civilized	0.73	Positive
honest	0.75	Positive
independent	0.80	Positive
creative	0.81	Positive
planned	0.82	Positive
logical	0.82	Positive
resolute	0.86	Positive
relaxed	0.87	Positive
original	0.87	Positive
patient	0.89	Positive
enterprising	0.91	Positive
reasonable	0.91	Positive
reliable	0.92	Positive
thorough	0.92	Positive
kind	0.92	Positive
strong	0.92	Positive
jolly	0.93	Positive
innovative	0.93	Positive
loyal	0.93	Positive

Overall Responses

Attribute	**Proportion of Desirable Response**	**Categorization**
unkind	0.17	Undesirable
dissatisfied	0.19	Undesirable
irresponsible	0.22	Undesirable
unfriendly	0.22	Undesirable
risky	0.29	Undesirable
fickle	0.29	Undesirable
intolerant	0.30	Undesirable
sarcastic	0.30	Undesirable
vain	0.30	Undesirable
fragile	0.30	Undesirable

lazy	0.33	Undesirable
vindictive	0.33	Undesirable
complicated	0.33	Undesirable
rigid	0.34	Undesirable
unreliable	0.35	Undesirable
impulsive	0.35	Undesirable
irritable	0.36	Undesirable
moody	0.37	Undesirable
frivolous	0.39	Undesirable
inhibited	0.40	Undesirable
sensitive	0.42	Undesirable
submissive	0.43	Undesirable
emotional	0.45	Undesirable
conformist	0.46	Undesirable
tempted	0.46	Undesirable
instinctive	0.46	Undesirable
indulgent	0.47	Undesirable
changeable	0.48	Undesirable
hedonistic	0.49	Undesirable
tough	0.52	Desirable
resistant	0.54	Desirable
certain	0.60	Desirable
rational	0.66	Desirable
logical	0.67	Desirable
methodical	0.67	Desirable
quick	0.68	Desirable
disciplined	0.68	Desirable
dependable	0.69	Desirable
planned	0.72	Desirable
industrious	0.73	Desirable
civilized	0.74	Desirable
independent	0.75	Desirable
determined	0.76	Desirable
relaxed	0.76	Desirable
loyal	0.77	Desirable
wise	0.78	Desirable
thorough	0.79	Desirable
original	0.82	Desirable
enterprising	0.83	Desirable
strong	0.83	Desirable
creative	0.83	Desirable
resolute	0.84	Desirable

patient	0.86	Desirable
reasonable	0.87	Desirable
jolly	0.89	Desirable
kind	0.90	Desirable
honest	0.90	Desirable
reliable	0.92	Desirable
innovative	0.95	Desirable

Responses by Male Raters

Attributes	**Proportion of Desirable Response**	**Categorization**
unfriendly	0.17	Undesirable
unkind	0.22	Undesirable
irresponsible	0.22	Undesirable
dissatisfied	0.26	Undesirable
vain	0.26	Undesirable
risky	0.29	Undesirable
sarcastic	0.29	Undesirable
fragile	0.31	Undesirable
lazy	0.32	Undesirable
vindictive	0.33	Undesirable
fickle	0.34	Undesirable
rigid	0.34	Undesirable
intolerant	0.35	Undesirable
complicated	0.35	Undesirable
impulsive	0.36	Undesirable
irritable	0.36	Undesirable
unreliable	0.41	Undesirable
moody	0.41	Undesirable
frivolous	0.43	Undesirable
submissive	0.47	Undesirable
inhibited	0.47	Undesirable
indulgent	0.47	Undesirable
emotional	0.48	Undesirable
hedonistic	0.48	Undesirable
sensitive	0.49	Undesirable
conformist	0.49	Undesirable
changeable	0.49	Undesirable
instinctive	0.52	Undesirable
tempted	0.52	Undesirable

tough	0.52	Desirable
resistant	0.53	Desirable
certain	0.61	Desirable
methodical	0.64	Desirable
rational	0.65	Desirable
industrious	0.68	Desirable
logical	0.69	Desirable
quick	0.71	Desirable
civilized	0.71	Desirable
disciplined	0.71	Desirable
relaxed	0.72	Desirable
planned	0.73	Desirable
loyal	0.74	Desirable
dependable	0.77	Desirable
determined	0.77	Desirable
independent	0.77	Desirable
original	0.82	Desirable
enterprising	0.83	Desirable
strong	0.83	Desirable
creative	0.84	Desirable
jolly	0.84	Desirable
wise	0.84	Desirable
resolute	0.85	Desirable
thorough	0.86	Desirable
kind	0.88	Desirable
honest	0.88	Desirable
patient	0.90	Desirable
reasonable	0.92	Desirable
innovative	0.97	Desirable
reliable	0.97	Desirable

Responses by Female Raters

Attributes	**Proportion of Desirable Response**	**Categorization**
unkind	0.11	Undesirable
dissatisfied	0.12	Undesirable
irresponsible	0.21	Undesirable
fickle	0.24	Undesirable
intolerant	0.25	Undesirable

unfriendly	0.27	Undesirable
risky	0.29	Undesirable
unreliable	0.29	Undesirable
fragile	0.29	Undesirable
sarcastic	0.31	Undesirable
complicated	0.31	Undesirable
vindictive	0.32	Undesirable
vain	0.33	Undesirable
lazy	0.33	Undesirable
rigid	0.33	Undesirable
impulsive	0.33	Undesirable
moody	0.33	Undesirable
inhibited	0.33	Undesirable
frivolous	0.35	Undesirable
sensitive	0.35	Undesirable
irritable	0.36	Undesirable
submissive	0.39	Undesirable
instinctive	0.40	Undesirable
tempted	0.40	Undesirable
emotional	0.43	Undesirable
conformist	0.43	Undesirable
indulgent	0.47	Undesirable
changeable	0.47	Undesirable
hedonistic	0.50	Undesirable
tough	0.52	Desirable
resisted	0.55	Desirable
certain	0.58	Desirable
dependable	0.60	Desirable
logical	0.64	Desirable
quick	0.65	Desirable
disciplined	0.65	Desirable
rational	0.67	Desirable
methodical	0.69	Desirable
planned	0.71	Desirable
thorough	0.72	Desirable
wise	0.72	Desirable
independent	0.73	Desirable
determined	0.75	Desirable
civilized	0.77	Desirable
industrious	0.78	Desirable

relaxed	0.80	Desirable
loyal	0.80	Desirable
patience	0.82	Desirable
reasonable	0.82	Desirable
original	0.82	Desirable
creative	0.82	Desirable
strong	0.83	Desirable
enterprising	0.83	Desirable
resolute	0.83	Desirable
reliable	0.87	Desirable
kind	0.92	Desirable
honest	0.92	Desirable
jolly	0.93	Desirable
innovative	0.93	Desirable

Appendix S4: Gender Bias by Attribute

Below, we present the average female (vs. male) bias associated with each attribute in the Common Crawl corpus. We calculated the difference in average similarity with female target words versus with male target words, (i.e., more positive numbers indicate greater relative similarity with female target words, and negative numbers indicate greater relative similarity with male target words). The positive attributes appear in the left column and negative attributes in the right column. Gender bias is exhibited when negative attributes are more strongly associated with female target words (i.e., positive numbers in the right column) and when positive attributes are more strongly associated with male target words (i.e., negative numbers in the left column).

Attribute	Average Bias
certain	-0.061
civilized	-0.048
creative	-0.004
dependable	-0.026
determined	-0.023
disciplined	-0.050
enterprising	-0.049
honest	-0.043
independent	0.010
industrious	-0.043
innovative	0.007
jolly	-0.049
kind	-0.058
logical	-0.052
loyal	-0.052
methodical	-0.059
original	-0.053
patience	-0.067
planned	-0.031
quick	-0.060
rational	-0.042
reasonable	-0.038
relaxed	0.013
reliable	-0.032
resisted	-0.046
resolute	-0.026
strong	-0.051
thorough	-0.021
tough	-0.072
wise	-0.080

Attribute	Average Bias
changeable	0.034
complicated	0.001
conformist	0.002
dissatisfied	0.008
emotional	0.018
fickle	0.008
fragile	0.050
frivolous	0.037
hedonistic	0.001
impulsive	0.003
indulgent	0.018
inhibited	0.052
instinctive	-0.028
intolerant	-0.008
irresponsible	-0.028
irritable	0.023
lazy	-0.033
moody	0.008
rigid	0.013
risky	-0.030
sarcastic	-0.005
sensitive	0.038
submissive	0.077
tempted	-0.042
unfriendly	-0.016
unkind	-0.008
unreliable	-0.033
vain	-0.047
vindictive	0.008

Appendix S5: Embeddings Analysis with Amazon Corpora

Extending our analysis from the primary text analysis study based on the Common Crawl text corpus to marketplace text corpora, we considered a product review text corpus from Amazon containing 7.9 million DVD movie reviews spanning 15 years. The Amazon data were obtained from McAuley and Leskovec (2013). In this analysis, we created our own 200-dimensional word embeddings using the GloVe algorithm developed by Pennington et al. (2014). We subsequently evaluated whether the word-embedding algorithm learned to associate female target words with more positive or more negative psychographic attributes relative to male target words, using the bias calculation described in Equation 1.

Results

Marketplace gender bias in Amazon reviews. We observed significant evidence of gender bias ($d = 1.01$, $p = 0.007$) in the Amazon reviews dataset. Specifically, we found that male target words had greater semantic similarity with positive psychographic attributes (1.707) than did female target words (1.241), ($d = 0.904$, $p < .001$). In addition, female words had greater semantic similarity with negative psychographic attributes (0.769) than did male words (0.746) ($d = 0.09$, $p < .001$).

Examining a marketplace text corpora spanning 7.9 million customer reviews on Amazon, our findings confirmed that algorithms learn gender biases from marketplace-related content. The word-embedding algorithm learned to associate women less closely with positive psychographic attributes and instead more closely with negative psychographic attributes relative to men.

Appendix S6: Robustness Checks

Names Analysis. In the main text, we study the similarity between consumer psychographic attributes and male/female pronouns. Building on this, we conducted additional analyses on the Common Crawl corpus that instead operationalized male vs. female dictionaries using common names (many of the names were not present within the Google Books dataset). Following prior literature (e.g., Greenwald et al., 1998, 2003; Nosek et al., 2002; Weyant, 2005), we used female and male name dictionaries that represented both Caucasian and African American ethnic identities. It is important to note, however, that names can be linked to additional characteristics beyond gender (e.g., SES). The names are shown in lowercase below.

Female target names: Aisha, Ebony, Keisha, Latoya, Tanisha, Shanice, Tamika, Raven, Joan, Lisa, Sarah, Diana, Kate, Ann, Amy, Donna.

Male target names: Darnell, Hakim, Jermaine, Kareem, Jamal, Leroy, Tyrone, Andre, John, Paul, Mike, Kevin, Steve, Greg, Jeff, Bill.

As in the analysis presented within the paper, we test for gender bias by comparing the similarity of target names (e.g., John, Jamal, Tanisha, Lisa) with positive and negative psychographic attributes (e.g., loyal, lazy, dissatisfied, rational).

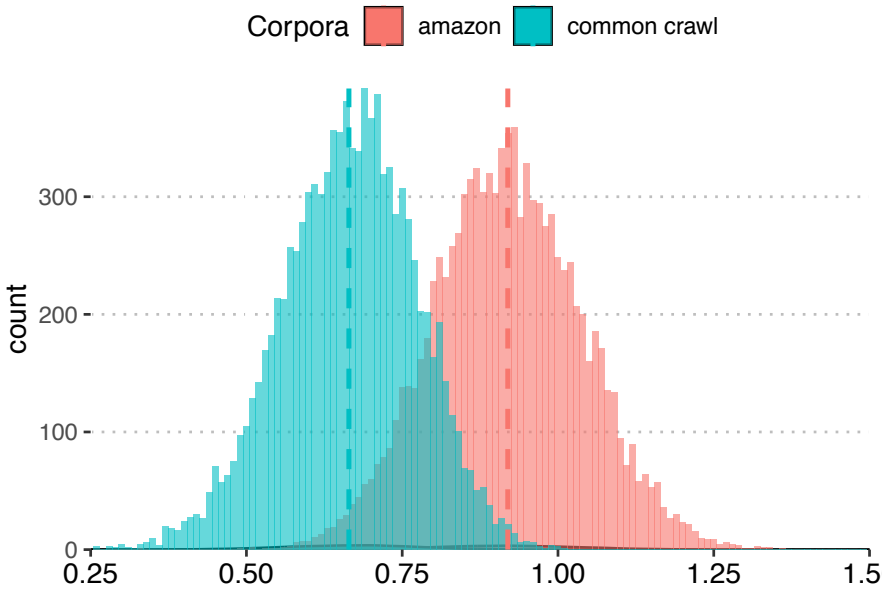
Common Crawl. The results confirmed that algorithms learned gender-biased consumer representations from Common Crawl ($d = 0.867$, $p = 0.0046$). Specifically, when we considered what drives gender bias, we found that male names were more closely associated (i.e., had a higher similarity) with positive attributes (0.340) than did female names (0.102), ($d = 0.213$, $p < .001$). However, female names had a higher similarity with negative attributes (1.027) than male names (0.543) ($d = 0.892$, $p < .001$).

Sensitivity Analysis. Our original marketplace attribute dictionary was generated by compiling psychographic attributes that have been studied in prior work in the literature. We conducted additional analysis to assess the degree to which the conclusions within the manuscript were sensitive to changes in the marketplace attribute dictionary. To this end, we generated an expanded list of more than 40 positive and 40 negative words, shown below, which ensured the inclusion of bipolar terms and additional attributes that could be construed as marketplace relevant (e.g., aggressive).

The expanded marketplace attribute dictionary included: honest–dishonest, reasonable–unreasonable, independent–dependent, thorough–careless, dependable–irresponsible, rational–emotional, relaxed–moody, loyal–fickle, reliable–unreliable, disciplined–disorderly, patience–impulsive, creative–uncreative, innovative–unimaginative, planned–instinctive, resolute–submissive, resisted–tempted, industrious–lazy, certain–uncertain, determined–unambitious, wise–foolish, tough–sensitive, jolly–irritable, civilized–uncivilized, strong–fragile, enterprising–unenterprising, quick–inhibited, logical–illogical, original–conformist, methodical–complicated, kind–unkind, friendly–unfriendly, flexible–rigid, tolerant–intolerant, prudent–indulgent, satisfied–dissatisfied, modest–vain, careful–risky, frugal–frivolous, forgiving–vindictive, consistent–changeable, frank–sarcastic, practical–hedonistic, amicable–aggressive.

We found an identical pattern of results using this expanded dictionary. In addition, in order to perform the sensitivity analysis, we generated 10,000 samples in which 30 positive and 30 negative words were randomly selected to create a new

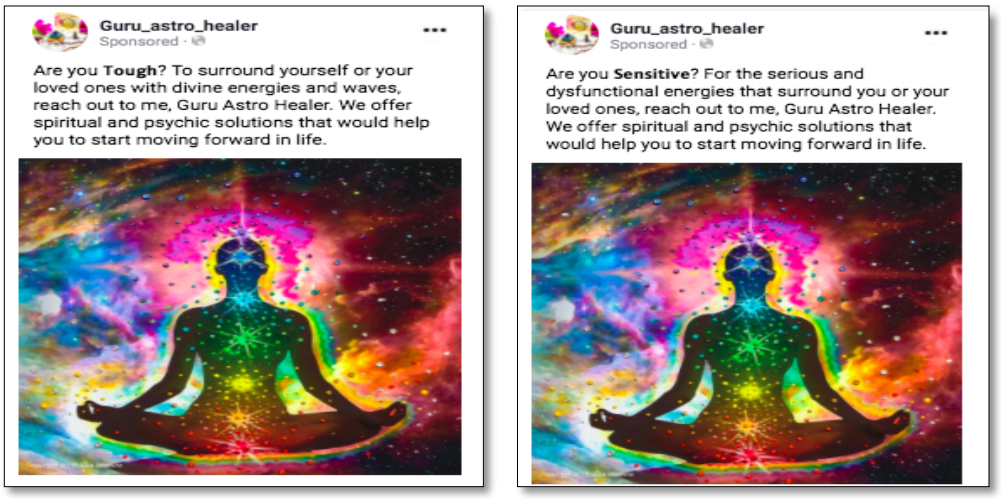
dictionary. For each sample, we then estimated the effect size, using the same procedures as those reported within the manuscript. Furthermore, we also examined gender bias using additional text corpora based on embeddings trained on millions of Amazon reviews. The histogram below shows effect size values for the 10,000 samples; recall that positive effect size shows bias against women (reported as *d* in the manuscript). This analysis confirmed that the gender bias we document in the manuscript is in fact a remarkably robust effect. Across samples and across corpora, we found evidence for large gender bias.



Appendix S7: Field Experiment – Gender Bias in Ad Targeting

Our word-embedding analyses demonstrated that algorithms learn gender-biased customer psychographic associations from stereotypes implicitly embedded within large text corpora. To evaluate consequences of this, we studied whether ad-targeting platforms, which leverage these algorithms, deliver biased product offerings to consumers in a manner consistent with the learned gender biases associated with the psychographic attributes. To do so, we partnered with an existing company to launch targeted ads on a major digital advertising platform.

Example Display Advertisement Stimuli



We collaborated with an astrologer who had been in practice for over two decades in the United States and wanted to leverage online advertising to reach a wider target audience. Partnering with the business, we developed a series of advertisements that varied the psychographic attribute used for targeting and compared positive attributes vs. negative attributes. We predicted that if algorithms have learned gender-biased consumer representations, they will consider ads targeting positive attributes to be more relevant to men vs. women and thus deliver positive ads to a greater share of men compared to ads targeting negative attributes. Thus, the "subject" in these causal tests was the ad-targeting algorithm and how it behaved (in terms of delivering ads to men and women).

Our manipulation, therefore, involved creating ten versions of the same advertisement, five including positive psychographic attributes and five including negative psychographic attributes. The five positive and five negative attributes shortlisted by the business were chosen to facilitate bipolar comparisons (strong–fragile, relaxed–moody, jolly–irritable, dependable–irresponsible, tough–sensitive). As the dependent measure, we observed the gender distribution of the consumers who were delivered the ad by the ad-targeting algorithm.

These advertisements were deployed on a major advertising platform (Facebook) as versions of the same campaign, with both versions deployed simultaneously. We chose "brand awareness" as the campaign objective; keywords used to select the audience were *healing*, *astrology*, and *chakra healing*; the bidding strategy was set to maximize

impressions, the location was limited to the United States, a specific gender was not targeted, and the campaign ran for five days.

Gender Distribution by Attribute

	Women	**Men**		**Women**	**Men**	χ^2	p
Strong	270 (9.58%)	2523 (88.28%)	Fragile	504 (19.03%)	2073 (78.29%)	106.78	< .001
Relaxed	250 (9.25%)	2390 (88.42%)	Moody	605 (21.61%)	2122 (75.81%)	162.94	< .001
Jolly	275 (9.63%)	2513 (88.02%)	Irritable	537 (21.90%)	1856 (75.42%)	156.75	< .001
Dependable	252 (9.60%)	2330 (88.80%)	Irresponsible	567 (22.22%)	1912 (74.95%)	169.45	< .001
Tough	279 (8.68%)	2861 (89.04%)	Sensitive	584 (21.71%)	2025 (75.39%)	204.82	< .001

Results

Over the duration of the ad campaign, the ads garnered a total of 26,720 impressions. To assess gender bias in ad delivery, we compared how ads targeting positive attributes vs. ads targeting negative attributes influenced the outcome of interest: the gender distribution of users receiving the ad.

Consistent with predictions, the valence of the psychographic attribute had a significant effect on the gender distribution of users considered relevant by the ad-targeting algorithm. Overall, across ads highlighting positive attributes, the ad-targeting algorithm delivered ads to users with a gender distribution of 9.5% women and 90.5% men. By comparison, across ads highlighting negative attributes, the algorithm delivered the ads to a significantly greater percentage of women (21.9%; 2797), and a lower percentage of men (78.1%; 9981), $\chi^2(1, N = 26{,}720) = 792.1$, $p < .001$. We note that the majority of users were male, similar to prior findings documented in the literature that suggest female users are more expensive to reach (Lambrecht & Tucker, 2019; Saez-Trumper et al., 2014). Nevertheless, when simply varying the psychographic attribute in the advertisement, we observed significant differences in the resulting gender distribution of users receiving the ads.

Furthermore, we found that the ad-targeting algorithm displayed gender bias within each of the attribute pairs. Women were significantly more likely than men to be targets of ads for those who are fragile vs. strong (19.0% vs. 9.5%), moody vs. relaxed (21.6% vs. 9.3%), irritable vs. jolly (21.9% vs. 9.6%), irresponsible vs. dependable (22.2% vs. 9.6%), and sensitive vs. tough (21.7% vs. 8.7%; all ps < .001). See Table 6. In line with Study 2, positive-attribute ads were considered to be more relevant to men vs. women.

The findings of this study are consistent with the idea that ad-targeting algorithms associate men with comparatively more positive psychographic attributes. While there is no scientific reason to believe that women are any less strong, relaxed, jolly, tough, or dependable than men (Hyde, 2016; Zell et al., 2015), ad-targeting algorithms on multiple major ad platforms consistently displayed this gender bias. We conducted a series of

follow-up studies to further characterize and establish the robustness of this finding. Specifically, we found similar results in a financial services domain and on another ad platform (Google; see appendices H and I). In addition, we demonstrate a way in which firms themselves can implement a debiasing strategy (Appendix S12).

Appendix S8: Field Experiment – Bias in Ads in the Investment Domain

This study focused on the financial services domain, where we simultaneously launched two ads in a between-participants design. The positive ad included all the positive psychographic attributes together (planned, disciplined, and creative), and the negative ad included all the negative psychographic attributes together (impulsive, dissatisfied, and irresponsible). For instance, users were presented with an ad that read "Save money for a better life: Investing tips for the planned, disciplined, and creative investor" or *"*Save money for a better life: Investing tips for the impulsive, dissatisfied, and irresponsible investor."

Because this study focused on the financial services domain, we created and hosted a website that offered readers basic tips on investing, which served as the destination page. The advertisement directed those who clicked on it toward the landing page, where they were able to learn basic investment strategies. Our manipulation involved creating two versions of the same advertisement, such that one included positive psychographic attribute words, and the second version included negative psychographic attribute words, hereafter referred to as the positive and negative ad, respectively. The positive ad said, "Save money for a better life: Investing tips for the planned, disciplined, creative investor." Whereas the negative ad said, "Save money for a better life: Investing tips for the impulsive, dissatisfied, irresponsible investor."

The two versions of the advertisement were identical in all respects except for the psychographic attribute words used. We also conducted a pretest (details presented below), which confirmed that the ads did not differ on either likability or realism, but were seen as more positive or more negative. We deployed these advertisements, released simultaneously, as versions of the same campaign on a large online advertising platform. We did not target a specific group of individuals, but we limited the location to the United States. We also selected the option of not giving preference to any particular gender so that from our end the advertising platform was instructed to show both types of advertisements with an equal likelihood to men and women.

Finally, the specific settings of this campaign were the same across both conditions: keywords were suggested by the ad platform, the bidding strategy was set to maximize clicks, and the campaign ran for six days. Settings were applied at the campaign level and, therefore, were identical for both positive and negative versions of the ad. Since the main aim of the study was to examine whether women (rather than men) would be targeted with advertisements containing negative or positive attribute words, our primary dependent variable was impressions (i.e., the number of times an advertisement was shown to users). We used the number of impressions for male and female consumers for each advertisement to evaluate our hypothesis.

Results

In total, across both positive and negative advertisements, the ad-targeting algorithm served 11,260 ad impressions. Crucially, we observed that positive vs. negative advertisements elicited a significant difference in the gender distribution of those who were served by the ad-targeting algorithm. Specifically, in the positive ad condition, the financial services advertisement was shown to consumers with a gender distribution of 13.3% (872) women and 86.7% (5670) men. However, when the ads included negative

psychographic attributes, the ad targeting algorithm delivered the ad to a significantly greater percentage of women (19.9%, 937) and a lower percentage of men (80.1%, 3781), $\chi^2(1, N = 11{,}260) = 86.2$, $p < .001$.

These findings indicate that the ad-targeting algorithm associated female users with negative psychographic attributes, and as a result woman were comparatively more likely to receive ads for companies targeting "impulsive, dissatisfied, irresponsible investors" than those targeting "planned, disciplined, creative investors." Finally, consistent with past work on clickthrough rates, the overall number of clickthroughs was low (155 clickthroughs across both advertisements). We did not observe any statistical difference ($\chi^2 < 1$) in clicking behavior across positive and negative advertisements between men and women, indicating that both sets of advertisements were considered relatively equal in terms of attention garnered.

Pretest for Ads in Investment Domain

Ninety-two participants from Amazon Mechanical Turk (MTurk) completed a pretest. The pretest used a (positive vs. negative) between-participants design. The participants were asked to rate how negative or positive (1 = "extremely negative" and 7 = "extremely positive") and how desirable (1 = "extremely undesirable," and 7 = "extremely desirable") they perceived the advertisement used in Study 1. They were also asked how much they liked and rated the advertisement (1 = "extremely dislike," and 7 = "extremely like") and if they would rate the advertisement as "real" or "not real." The results showed that participants perceived the positive advertisement as more positive ($M_{positive} = 5.40$, $M_{negative} = 4.71$; $F(1, 91) = 4.31$, $p = .04$) and more desirable ($M_{positive} = 5.36$, $M_{negative} = 4.76$; $F(1, 91) = 3.38$, $p = .07$) than the negative advertisement. On the other hand, there was no difference in the likability ($M_{positive} = 5.06$, $M_{negative} = 5$; $F(1, 91) = 0.03$, $p = .86$) or realism ($P_{positive} = 85\%$, $P_{negative} = 76\%$; $F(1, 91) = 1.33$, $p = .25$) in their evaluation of the positive versus negative advertisement.

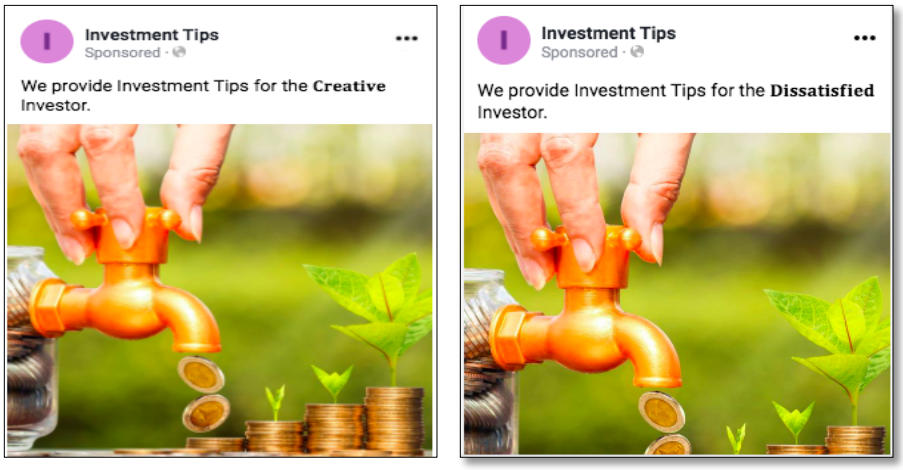
Appendix S9: Field Experiment – Bias in Ads in Investment Domain, Replication

This study replicated the findings from the previous study (Appendix S8) on another major ad platform. We followed the same general approach as in the previous study, focusing on the financial services domain; however we modified the design in a few ways to build on the previous study. First, we analyzed each psychographic attribute individually by including a single attribute in each advertisement, such that the manipulation involved only a single word change in the ad copy. Second, to gain clearer insight into algorithmic bias, we applied an impressions-optimization goal limiting the degree of user feedback that the ad targeting algorithm could adapt.

Our manipulation involved creating six versions of the same advertisement, applying a single attribute in each ad. Three versions included positive psychographic attributes (*planned, disciplined*, and *creative*), and the other three versions included negative psychographic attributes (*impulsive, dissatisfied*, and *irresponsible*)—hereafter referred to as the positive and negative ads, respectively (see the figure below for examples of the positive and negative versions). For instance, users were presented with an ad that read "Save money for a better life: Investing tips for the planned investor" or "Save money for a better life: Investing tips for the impulsive investor."

We simultaneously deployed all six versions of the advertisement as part of the same campaign on a major advertising platform (Facebook). We chose "brand awareness" as the campaign objective, and the keywords were set as *investment strategy*, *daily investment tips*, and *investor*; the bidding strategy was set to maximize impressions, the location was limited to the United States, and the campaign ran for three days. Finally, we chose the option of not giving preference to any particular gender, such that from our end the advertising platform was instructed to show both types of advertisements with an equal likelihood to men and women. Impressions served as the dependent variable to examine whether ad-targeting algorithms served women with more negative or positive advertisements.

Example Display Advertisement Stimuli



Results

Overall, 18,729 impressions were delivered by the ad-targeting algorithm. We

found significant differences in the gender distribution of users targeted by positive-attribute versus negative-attribute ads. The targeting algorithm served the positive-attribute ads to users with a gender distribution of 8.8% (817) women and 89.1% (8281) men. In comparison, the targeting algorithm served the negative ads to significantly more women (10.8%, 1014) and fewer men (87.1%, 8220), $\chi^2(1, N = 18{,}332) = 20.4$, $p < .001$. Please see the table below for a breakdown of the gender distribution for each attribute.

Consistent with our predictions, these findings show that relative to positive ads, negative ads were comparatively more likely to be delivered to women than to men. This result suggests that ad-targeting algorithms may consider women to be relevant to more negative vs. positive psychographic attributes compared to men. These findings provide initial evidence for gender-biased consumption bubbles: Ad campaigns targeting planned, disciplined, and creative investors were significantly more likely to be delivered to men compared to those ads targeting impulsive, dissatisfied, and irresponsible investors, which were more likely to be delivered to women.

Gender-Distribution by Attribute

Positive attributes	**N**	**Females**	**Males**
Planned	3197	285 (8.91%)	2841 (88.86%)
Disciplined	3288	278 (8.45%)	2936 (89.29%)
Creative	2811	254 (9.03%)	2504 (89.08%)

Negative attributes	**N**	**Females**	**Males**
Impulsive	3509	373 (10.63%)	3060 (87.20%)
Irresponsible	3425	394 (11.50%)	2956 (86.31%)
Dissatisfied	2499	247 (9.88%)	2204 (88.20%)

Appendix S10: Study 2 Supplementary Methods and Results

Study 2

Methods. Participants were asked to create a new account even if they already had one and use the account while browsing for one day. Creating a new account helped to minimize any influence of shared browsing history that might occur when the same computing device is shared with many people; it also helped ensure that the shopping portal could learn the gender of the user (while Google asks the gender of the account holder explicitly and quite accurately guesses the gender when not available, Bing learns the gender through search behavior; Duhaime-Ross 2014).

Mediation analysis. To evaluate how the consideration set influenced the actual choice of the product, we conducted a mediation analysis. Because our data have a multilevel structure, we tested this hypothesis following the procedure outlined by Tingley et al. (2014) and implemented in the R package "mediation." Following this process, two random intercept models were fit. The first model had consideration-set ratings as the outcome variable, while gender of the participant who saw the screenshot, source, and rater's gender were entered into the model as fixed effects. The second model had choice ratings as the outcome variable, while participant's gender, source, rater's gender, and consideration-set ratings were entered as fixed effects. These fitted models were used in the mediation analysis. To estimate the confidence interval around the treatment effect, direct effect, and average total effect, we performed 1000 simulations using the quasi-Bayesian Monte Carlo method based on normal approximation (Imai, Keele, & Yamamoto, 2010). We found that the average total effect was significant (β = .149, p = .02, 95%CI = [.02,0.27]), suggesting that females selected products with negative attributes more often than males (a higher score means a more negative attribute). When we decomposed this effect into direct and indirect effects, we found that the portion of the average total effect that was transmitted through the consideration set (i.e., the indirect effect) was statistically significant (β = .134, p < .001, 95%CI = [0.06, 0.21]). This suggests that the women's choice of a negative product was influenced by the negative consideration set shown to them. The direct effect (excluding the mediator) was not significant (β = .015, p =.76, 95%CI = [-.08, .11]). Overall, this analysis suggests that the biased consideration set facilitated the final choice of product.

Appendix S11: Study 3 Supplementary Results

Study 3

Below we report the number of impressions and clickthroughs for both male and female users in each condition of the 2 (ad targeting attribute: negative vs. positive psychographic attribute) × 2 (campaign objective: impressions optimization vs. clickthrough optimization) study.

Impressions and Clickthroughs by User Gender

Impressions

Attribute	User Gender	Clickthrough Optimization	Impressions Optimization
Positive (*strong*)	Women	2129 (3885); 54.8%	1644 (4500); 36.5%
Positive (*strong*)	Men	1756 (3885); 45.2%	2856 (4500); 63.5%
Negative (*fragile*)	Women	2522 (3578); 70.5%	1679 (4319); 38.9%
Negative (*fragile*)	Men	1056 (3578); 29.5%	2640 (4319); 61.1%

Clickthroughs

Attribute	User Gender	Clickthrough Optimization	Impressions Optimization
Positive (*strong*)	Women	23 clicks	2 clicks
Positive (*strong*)	Men	9 clicks	1 click
Negative (*fragile*)	Women	41 clicks	1 click
Negative (*fragile*)	Men	8 clicks	2 clicks

Analysis within clickthrough-optimization only. In Figure 3 in the main text, we plot the percentage of ads delivered to women on the Y-axis separately for negative-attribute and positive-attribute ads. Focusing on the clickthrough-optimization condition only, we find that at the launch of the ad campaign (during which there is little user input) algorithms are biased in their delivery of negative-attribute ads more frequently to women (54.7% vs. 49.0% at time=1). The gender bias is magnified over time such that a greater bias is observed at time=40 (70.5% vs. 54.8%), reflecting co-production of the bias through user acceptance of gender stereotypes. Linear regression analysis of percent of women ad recipients on attribute valence, timepoint, and their interaction revealed a significant interaction effect (b = .039%, se = .12%, $t(76) = 3.44$, p = .001), supporting the conclusion that gender bias is co-produced by algorithms and users.

Time-series analysis. The data were recorded over 40 consecutive, evenly spaced timepoints. Such datasets can display autocorrelation, and analyzing them with linear regression has the potential of violating the assumption of independence of residual. To address this concern, as the first step we ran the Durbin-Watson test to identify if autocorrelation is present in the residuals of the following linear regression model that incorporated all the two-way and three-way interactions among independent variables. Here T_t represents the time points, where t ranged from 1 to 40. P_j represents the

campaign objective, where the value of j was either 0 (impression optimization) or 1 (clickthrough optimization). A_k represents the type of ad, and k was either 0 (positive ad) or 1 (negative ad). The dependent variable y represents the fraction of women shown ad k on platform j at time point t.

$$y = \alpha + \beta_1 * T_t + \beta_2 * P_j + \beta_3 * A_k + \beta_4 * T_t * P_j + \beta_5 * T_t * A_k + \beta_6 * P_j * A_k + \beta_7 * T_t * P_j * A_k + \epsilon$$

Durbin-Watson Test

Lag	Autocorrelation	D-W Statistic	p-value
1	0.735871234	0.5271669	0.000
2	0.528638291	0.9414842	0.000
3	0.309924171	1.3788171	0.000
4	0.093301828	1.8093277	0.152
5	0.008217387	1.9792800	0.784

The Durbin-Watson test showed that for at least the first three lags there was significant autocorrelation. To incorporate the presence of residual autocorrelation, we used generalized linear models with various specifications of autoregressive (AR) and moving average (MA) processes. Starting with the first-order autoregressive and first-order moving average process for residuals, ARMA(1, 1), we ran three additional models with ARMA(2, 2), ARMA(3, 3), and ARMA (4, 4) process. We then used likelihood-ratio tests to assess if a more complicated specification of ARMA process is necessary, and whether a simpler model is sufficient. The results reported below support the ARMA(4, 4) process.

ARMA Processes	No ARMA	ARMA (1,1)	ARMA (2,2)	ARMA (3,3)	ARMA (4,4)
No ARMA specifications AIC = -1039.099		137.141 p <.0001	144.82029 p <.0001	149.26288 p <.0001	159.7195 p <.0001
ARMA (1,1) AIC = -1172.240			7.67894 p = 0.0215	12.12153 p = 0.0165	22.57815 p < 0.001
ARMA (2,2)				4.44259	14.89921

AIC = -1175.919				p = 0.1085	p = 0.0049
ARMA (3,3) AIC = -1176.361					10.45662 p = 0.0054
ARMA (4,4) AIC = -1182.818					

The estimates of the regression parameters under the ARMA(4, 4) error-correlation model are reported below.

	Value	**Standard Error**	***t*-value**	***p*-value**
(Intercept)	0.3746	0.00351	106.486	0.0000
Time	-0.00001	0.00015	-0.078	0.9377
Ad	0.01232	0.00450	2.736	0.0070
Platform	0.09709	0.00460	21.084	0.0000
Time x Ad	0.00005	0.00019	0.254	0.7991
Time x Platform	0.00245	0.00020	11.880	0.0000
Ad x Platform	0.08485	0.00629	13.489	0.0000
Time x Ad x Platform	0.00095	0.00028	3.325	0.0011

Appendix S12: Field Experiment – Debiasing Strategies

This study builds a bridge between the embedding studies and the ad-targeting studies to illustrate how bias strength in word embeddings is related to the degree of bias in ad delivery. To this end, we focused on positive ads only and compared strong vs. weak gender-biased psychographic attributes, helping to illustrate how gender bias learned from text corpora and applied within advertisements as attributes used for targeting can result in gender-biased ad delivery. Moreover, this study serves to demonstrate a way in which firms themselves can approach debiasing. Large ad platforms typically use proprietary and opaque algorithms, while firms that use these platforms to deploy their advertisements have no say in the design of the platform's algorithms. The current study illustrates that by looking at word embeddings, firms can design ad copy to include words that can reduce gender bias in ad delivery.

As in the study in Appendix S7, we again collaborated with an astrology firm to run their ads on a major ad platform. We followed the same general procedures as in previous studies, where we manipulated the attributes in the advertisement and observed the resulting gender distribution. The design included three different advertisements: a strong-positive condition, a weak-positive condition, and a no-attribute control. We used the word embeddings from the text analytics study to identify strong-positive and weak-positive attributes (i.e., a positive psychographic attribute that had strong differential association with men vs. women rather than one that had weak differential association). In particular, we selected the strong-positive attribute *tough* (average bias = -0.072, Appendix S4) and the weak-positive attribute *creative* (average bias = -0.004). If an attribute word were equally associated with men and women in the text corpus, algorithms learning from the corpus would learn to associate the attribute as much with women as with men, so there would be no gender-biased association learned. Thus, when these bias-neutral ads are used, they should result in an equal number of men and women being shown the ads compared to baseline. However, if a positive attribute is strongly associated with men vs. women, then it would result in more ads containing that word shown to men compared to baseline.

Serving as the baseline, the no-attribute control ad read: "Calling all people! Discover your mind, soul, and consciousness." The strong-positive and weak-positive ads, respectively, read: "Calling all tough/creative people! Discover your mind, soul, and consciousness." These advertisements were deployed on Facebook as versions of the same campaign, with all the three versions deployed simultaneously. We chose "brand awareness" as the campaign objective; keywords used to select the audience were again *healing, astrology*, and *chakra healing*. The bidding strategy was set to maximize impressions, the location was limited to the United States, and a specific gender was not targeted. The budget was set as a lifetime budget for the campaign, which ran for one day. We examined the resulting gender distribution as the dependent variable for our analysis.

Results

In total there were 30,015 impressions, and men were shown the majority of the advertisements (21,777) by the ad-targeting algorithm. Important to note, we found that manipulating the psychographic attribute words had a significant influence on the

resulting gender distribution of users who were delivered the ads by applying the Tukey method to test contrasts between the three conditions. Consistent with previous findings, we again found that the ad-targeting algorithm delivered ads with strong-positive attributes significantly more often to men vs. women relative to the no-attribute control (strong-positive: 22.6% women, 75.2% men; no-attribute control: 26.6% women, 71.6% men; $t(20{,}036) = 6.39$, $p < .001$). However, the firm was able to eliminate this gender bias in ad delivery by applying a weak-positive attribute in the ad. Specifically, the ad-targeting algorithm delivered the weak-positive advertisement to users (26.9% women, 70.9% men) with a similar gender distribution as the no-attribute control advertisement ($t(20{,}266) = 0.52$, $p = .86$).

Gender-Distribution by Attribute

Attribute	N	Females	Males
No-Attribute Control	10287	2732 (26.56%)	7364 (71.59%)
Strong-Positive: Tough	9749	2206 (22.62%)	7330 (75.18%)
Weak-Positive: Creative	9979	2682 (26.88%)	7083 (70.98%)

Therefore, this study indicates that selection of attributes that are strongly or weakly associated with men and women (identified through word embeddings) can help reduce prejudicial or preferential ad delivery. This empowers firms to make their ad delivery more equitable and offers them a practical debiasing strategy prior to deploying ads on large ad platforms.