Supplementary Information

for

Hiring women into senior leadership positions is associated with a reduction in gender stereotypes in organizational language

M. Asher Lawson*

Duke University

Ashley E. Martin

Stanford University

Imrul Huda

University of Chicago

Sandra C. Matz

Columbia University

*Corresponding author: M. Asher Lawson

Email: asher.lawson@duke.edu

Table of Contents

Selection of target and propensity-matched firms

We present the 11 target firms that fulfilled our selection criteria, along with their CEOs' names and takeover dates below in Table S1a. Additionally, we include two companies that went from having a female CEO to a male CEO within the target period (Table 1b), subject to the same criteria as the male-female hires (i.e., at least three years of data available, no spin offs, and of course, no male predecessors).

Table S1a: Target companies and associated female CEOs with hiring dates.

director name	takeover	company name	TICKER
BARBARA RENTLER	6/1/14	ROSS STORES, INC.	ROST
HEATHER BRESCH	1/1/12	MYLAN NV	MYL
LISA SU	10/8/14	ADVANCED MICRO DEVICES, INC.	AMD
LYNN GOOD	7/1/13	DUKE ENERGY CORP.	DUK
MARILLYN HEWSON	1/1/13	LOCKHEED MARTIN CORP.	LMT
MARISSA A. MAYER	7/17/12	YAHOO! INC.	YHOO
MARY DILLON	7/1/13	ULTA BEAUTY, INC.	ULTA
MARY BARRA	1/15/14	GENERAL MOTORS CO.	GM
PHEBE NOVAKOVIC	1/1/13	GENERAL DYNAMICS CORP.	GD
SAFRA CATZ (CO-CEO)	9/18/14	ORACLE CORP.	ORCL
VIRGINIA ROMETTY	1/1/12	INTERNATIONAL BUSINESS MACHINES CORP.	IBM

Table S1b: Target companies and associated female CEOs who were replaced.

director name	takeover	company name	TICKER
ANGELA BRALY	8/28/12	WELLPOINT / ANTHEM	WLP / ANTM
PATRICIA WOERTZ	1/1/15	ARCHER DANIELS MIDLAND	ADM

In order to study the effects of appointing a female CEO on linguistic stereotypes, we implemented a propensity matching procedure to identify a group of organizations that were matched on relevant pretreatment characteristics (1). Rather than implementing a full propensity score matching algorithm, we theoretically reasoned the most important variables to balance between target organizations and their propensity matches, searched for matches based on these variables, and empirically verified that we did indeed successfully balance these characteristics.

One notable difficulty in comparing organizations based on pretreatment characteristics is the high-dimensional vector of characteristics by which they can be compared (2). One could end up trading off rather large differences in important variables for greater alignment across a diffuse set of variables. In order to avoid such dimensionality issues, we elected to search for propensity matches based on a concentrated set of the most relevant organizational characteristics.

Subject to the constraint of choosing a small set of highly relevant characteristics, we considered what dimensions of organizations were most relevant to our research question of interest. The propensity to hire a female CEO could be affected by many different factors. We specifically focused on organizations' size, age, and industry to control for effects associated with variation in company culture and attitudes towards women. Whilst it is possible that more proximal variables (such as stock price changes) may affect companies' propensity to replace their CEO, culture is more likely to determine the likelihood of appointing specifically a female CEO, and attitudes towards women more broadly. Small organizations may have more fluid cultures and norms that will be more likely to perform rare actions (such as hiring women as CEOs), and may respond more vigorously to changes in personnel than more established organizations. Similarly, new organizations may be less inertial than older ones, and thus exhibit different and less stable attitudes regarding gender that could affect the propensity to hire a female CEO. Finally, there are massive difference in the properties conducive for success across industry, and thus we might reasonably expect organizations in, for example, the tech industry versus the construction industry to select and characterize their leaders in different manners. For these reasons, we elected our focused set of three variables with which to select propensity matches.

Details of the procedure by which we searched for propensity matched organizations are included in the Methods section of the main manuscript. In sum, for each of the 11 target companies and the 2 female-male hire companies we identified two propensity-matched companies. To do so, we focused on three primary variables; the size of the company as measured by the number of employees, the age of the company (as measured by the years since its founding date), and company's industry code (SICS).

We provide two logistic regression models below indicating the success of our procedure – none of the variables attained statistical significance in predicting whether the organization was a target organization (that hired a female CEO) versus a propensity match. We estimated models both using the original set of 33 organizations (Model 1), and the later expanded set of 39 organizations (Model 2). We used the first digit of the SICs industry code as a proxy measure for the regression modelling (but in practice aimed to match on as many digits as possible). These model estimates can be seen in Table S2.

Table S2: Logistic regression models showing successful propensity matching procedure.

Variable	Model 1	Model 2
(Intercept)	1.053*	1.019†
	(0.487)	(0.501)
Number employees	0.000	0.000
	(0.000)	(0.000)
Years Old	-0.007	-0.003
	(0.005)	(0.004)
SIC (1)	-0.542	-0.521
	(0.563)	(0.579)
SIC (2)	-0.346	-0.520
	(0.559)	(0.552)

SIC (3)	-0.946†	-0.990†
	(0.550)	(0.565)
SIC (4)	-0.561	-0.625
	(0.531)	(0.543)
SIC (5)		-0.505
		(0.608)
SIC (6)	-0.464	-0.538
	(0.532)	(0.544)

See Table S3 for a full list of these companies, and for full details of our materials and registrations, please see here: https://osf.io/zfq3d and here https://osf.io/utz29/. We did not change our target companies at any point in our analyses, but did update the CIK (Central Index Key) code for Dollar Tree, after it came to our attention that the identifier was incorrect. The two female-male hires and their associated propensity matches were added later, but their propensity matches were also recovered prior to analyzing the text data.

Table S3: Target and propensity matched companies identified by CIK code.

Type	Firm number / propensity match	0		1		2		Date
Mas-Fem	1	ROSS STORES, INC.	0000745732	Dollar Tree, Inc.	0000935703	Kohl's Corporation	0000885639	6/1/14
Mas-Fem	2	MYLAN NV	0001623613	Amgen Inc.	0000318154	The Estee Lauder Companies Inc.	0001001250	1/1/12
Mas-Fem	3	ADVANCED MICRO DEVICES, INC.	0000002488	Skyworks Solutions, Inc.	0000004127	Broadridge Financial Solutions, Inc.	0001383312	10/8/14
Mas-Fem	4	DUKE ENERGY CORP.	0001326160	Exelon Corporation	0001109357	Republic Services, Inc.	0001060391	7/1/13
Mas-Fem	5	LOCKHEED MARTIN CORP.	0000936468	Thermo Fisher Scientific Inc.	0000097745	Danaher Corporation	0000313616	1/1/13
Mas-Fem	6	YAHOO! INC.	0001011006	Citrix Systems, Inc.	0000877890	Akamai Technologies, Inc.	0001086222	7/17/12
Mas-Fem	7	ULTA BEAUTY, INC.	0001403568	L Brands, Inc.	0000701985	YUM! Brands, Inc.	0001041061	7/1/13

Mas-Fem	8	GENERAL MOTORS CO.	0001467858	The Boeing Company	0000012927	Ford Motor Cmpany	0000037996	1/15/14
Mas-Fem	9	GENERAL DYNAMICS CORP.	0000040533	Honeywell International Inc.	0000773840	Cummins Inc.	0000026172	1/1/13
Mas-Fem	10	ORACLE CORP.	0001341439	Microsoft Corporation	0000789019	DXC Technology Company	0001688568	9/18/14
Mas-Fem	11	INTERNATIONAL BUSINESS MACHINES CORP.	0000051143	General Electric Company	0000040545	United Parcel Service, Inc.	0001090727	1/1/12
Fem-Mas	1	ANTHEM, INC.	0001156039	Humana Inc.	0000049071	The Allstate Corporation	0000899051	8/28/12
Fem-Mas	2	THE ARCHER-DANIELS-MIDLAND COMPANY	0000007084	Kellogg Company	0000055067	Bristol-Myers Squibb Company	0000014272	1/1/15

Dictionary development and validation

Agency and communality dictionaries. First, we conducted a comprehensive literature review of papers published on agentic and communal language (3–21). From these papers, we identified 356 agency-relevant and 278 communality-relevant words and phrases. After removal of phrases and duplicate items, we retained 221 agency words and 162 communality words.

In order to construct our agency and communality dictionaries, we collected ratings of the words on three dimensions; agency, communality, and valence. We prefaced the ratings with the following descriptions:

Agency definition:

An individual’s striving to be independent, control one's environment, and to assert, protect, and expand one's self. Agentic individuals are autonomous and individualistic. They strive to achieve their goals, experience achievement, and master their environment, even if they have to conquer obstacles and dominate others. Agency-oriented individuals experience fulfillment through their individual accomplishments and their sense of independence and separateness from others.

Communality definition:

A person’s striving to be part of a community, establish close relationships and connect with others. These individuals are empathetic and understanding. They strive to closely relate and cooperate and merge with others, even if they sometimes must sacrifice their individual needs for the common good. They experience fulfillment through their group accomplishments, close relationships, and a sense of belonging.

Valence definition:

Valence describes how positive or negative you think a word is.

Participants rated the words on a 7-point scale ranging from 1-7, with the following instructions:

Ratings:

Please rate items on the agency scale from 1-7, where '1' is equivalent to saying that you do not think this word captures the meaning of agency at all, and '7' means you think this word fully captures the meaning of agency.

Please rate items on the communality scale from 1-7, where '1' is equivalent to saying that you do not think this word captures the meaning of communality at all, and '7' means you think this word fully captures the meaning of communality.

Please rate items on the valence, where '-3' means you think this is a very negative word, and '+3' means it is a very positive word. '0' implies that you think the word is neutral.

Table S4a presents all the 100 words in the agency dictionary alongside the means and SDs of the agency and valence ratings. Some examples are plotted in agency-valence space in Figure S1. Table S4b contains the mean and SD of communality and valence ratings for the communality dictionary. Further, Table S4c contains inter-rater agreement statistics from the study where we collected ratings of each dictionary.

Inter-rater agreement was calculated using r_{wg} (22), which captures inter-rater reliability when a group of judges rates a single target on a single variable. The r_{wg} coefficient is impacted both by the empirical variance in the judges' ratings, and the estimated variance in the ratings if the judges' ratings were random. The agreement coefficient can be interpreted as the proportion of variance that is associated with agreement between judges (23). We computed r_{wg} using the "multilevel" package for R (24)

The average r_{wg} for an agency word being rated for agency was 0.546; the average r_{wg} for a communality word being rated for communality was 0.550. For valence, agency words had an average r_{wg} of 0.661 and communality words had an average of 0.709.

Table S4a: Agency and valence ratings of 100-word dictionary.

	Agency		Valence	
Word	M	SD	M	SD
achievement	6.11	1.24	1.87	1.02
active	5.65	1.33	1.75	0.96
adamant	5.76	1.38	0.29	1.29
aggressive	5.27	1.75	-1.17	1.45
ambition	6.31	1.06	1.38	1.25

ambitious	6.36	0.79	1.56	1.03
analytical	5.88	1.04	1.59	1.17
assert	5.98	1.28	0.70	1.23
assertive	6.04	1.35	0.90	1.36
assertiveness	6.06	1.31	0.78	1.42
assured	5.45	1.51	1.46	1.05
autonomous	6.23	1.29	0.96	1.33
autonomy	6.47	1.00	1.16	1.18
bold	5.90	1.36	1.35	1.16
bossy	5.35	1.91	-1.52	1.33
brave	5.98	1.18	2.18	1.09
brilliant	5.75	1.31	2.39	0.81
capability	5.94	1.31	1.92	1.11
capable	6.00	1.30	2.13	0.88
clever	5.73	1.47	1.81	1.12
command	5.90	1.32	0.45	1.31
competence	5.73	1.33	1.88	1.03
competent	5.82	1.36	2.06	0.97
competitive	6.12	1.39	0.10	1.28
competitiveness	6.18	1.36	0.13	1.44
confidence	6.06	1.28	1.98	1.03
confident	6.25	1.19	2.04	0.98
convincing	5.42	1.49	1.12	1.07
creative	5.63	1.42	2.08	1.01
cunning	5.24	1.78	-0.78	1.50
daring	5.67	1.48	0.86	1.14
decisive	6.00	1.21	1.44	1.03
determined	6.06	1.34	1.96	0.93
diligent	6.04	0.91	2.02	1.06
direct	5.79	1.44	1.06	1.23
dominant	5.81	1.55	-0.31	1.29
dynamic	5.56	1.17	1.67	0.93
educated	5.29	1.58	2.10	0.90
effective	5.79	1.40	2.00	0.91
efficient	5.92	1.26	1.98	0.90
egocentrist	5.44	1.74	-1.65	1.33
egoistic	5.43	1.73	-1.69	1.29
energetic	5.44	1.50	1.73	1.09
exploration	5.84	1.20	1.47	1.19
fast	5.22	1.79	1.00	1.08
freedom	6.38	0.85	1.88	1.12
hardworking	5.96	1.41	2.30	0.81
imaginative	5.46	1.24	1.73	1.12
independence	6.54	1.03	1.63	1.18
independent	6.72	0.66	1.93	0.93
individual	6.49	1.29	0.92	1.24
individualistic	6.42	1.20	0.58	1.23
industrious	5.73	1.22	1.38	1.41
ingenious	5.21	1.49	1.53	1.33
insightful	5.60	1.22	2.00	0.96
intellectual	5.53	1.37	1.90	1.15
intelligent	5.81	1.48	2.00	1.24
knowledgeable	5.69	1.17	2.08	1.00
leader	6.32	0.96	1.94	1.08
logical	5.59	1.51	1.86	1.05
meticulous	5.42	1.51	1.39	1.20
organized	5.75	1.28	2.04	0.89
original	5.83	1.26	1.45	1.19
outspoken	5.46	1.47	0.74	1.17
perceptive	5.27	1.59	1.86	1.10
persistent	5.98	1.05	1.34	1.04
power	5.98	1.13	0.66	1.39
powerful	5.90	1.28	1.10	1.28
practical	5.50	1.44	1.58	1.03

proud	5.40	1.62	0.85	1.29
rational	5.60	1.22	1.81	0.98
realist	5.52	1.60	0.98	1.02
recognition	5.50	1.41	0.96	1.14
resilient	5.72	1.39	1.91	1.04
resourceful	6.00	1.21	1.92	0.93
self-assertion	6.17	1.17	0.69	1.45
self-confident	6.40	0.89	1.60	1.21
self-contained	5.63	1.44	0.17	1.40
self-control	5.54	1.68	1.42	1.22
self-direction	6.39	0.95	1.27	1.18
self-expansion	5.83	1.49	0.61	1.32
self-important	5.58	1.71	-0.94	1.53
self-protection	6.08	1.22	0.65	1.42
self-reliant	6.59	0.76	1.56	0.99
self-sufficient	6.67	0.63	1.87	0.93
serious	5.51	1.25	0.68	1.18
sharp	5.56	1.41	1.51	1.18
skillful	5.75	1.39	2.10	0.97
skillfulness	5.81	1.26	2.23	0.89
smart	5.67	1.30	2.10	1.17
status	5.24	1.39	0.27	1.35
strong	5.98	1.42	2.17	0.91
strong-willed	6.33	1.13	1.08	1.40
superiority	5.35	1.71	-0.96	1.40
tough	5.33	1.42	0.62	1.44
unique	5.68	1.32	1.59	1.06
unwavering	5.41	1.33	0.80	1.28
vigorous	5.39	1.73	0.91	1.23
well-disciplined	6.00	1.29	1.82	1.01
well-organized	5.53	1.32	1.69	1.29

Table S4b: Communality and valence ratings of 100-word dictionary.

	Communality		Valence	
Word	Mean	SD	Mean	SD
accepting	6.04	1.30	1.90	1.06
affectionate	6.00	1.47	1.76	1.15
agreeable	5.78	1.52	1.57	1.12
altruism	5.92	1.30	1.61	1.47
altruistic	5.88	1.49	1.73	1.41
attachment	5.57	1.60	0.39	1.16
belonging	6.33	1.23	1.49	1.04
benevolence	5.37	1.78	1.50	1.34
care-taking	6.17	1.00	1.57	1.08
caring	6.16	1.33	1.94	0.95
cheerful	5.43	1.58	1.90	1.08
civility	6.06	0.89	1.62	0.90
closeness	5.92	1.41	1.34	1.26
communal	6.47	1.23	0.94	1.34
communicative	6.21	1.34	2.02	0.93
compassion	6.29	1.17	2.11	0.89
compassionate	6.23	1.21	2.14	1.10
compromising	5.64	1.61	0.75	1.42
connected	6.53	0.89	1.67	1.02
connections	6.26	1.23	1.41	0.96
conscientious	5.58	1.15	1.69	1.08
considerate	6.33	0.82	2.04	1.04
consideration	5.98	1.03	1.68	0.98
cooperation	6.63	0.67	1.96	0.91

cooperative	6.61	0.81	1.90	0.81
dependable	5.88	1.25	2.06	1.16
dependency	5.67	1.31	-0.46	1.49
dependent	5.60	1.50	-0.26	1.55
duty	5.41	1.61	1.21	1.18
easygoing	5.53	1.49	1.69	1.17
enthusiastic	5.35	1.54	1.98	0.86
equality	6.22	1.14	1.92	1.37
fair	5.59	1.31	1.88	1.00
fair-minded	5.54	1.61	1.50	1.37
faithful	5.92	1.34	1.71	1.07
family-oriented	6.26	1.17	1.73	1.12
flexible	5.60	1.57	1.63	0.95
forgiveness	5.88	1.30	1.78	0.98
forgiving	5.84	1.16	1.75	0.89
friendliness	6.16	1.28	1.90	1.02
friendly	6.27	1.14	2.20	0.84
generous	6.14	1.38	2.04	1.01
gentle	5.60	1.33	1.50	0.97
good	5.51	1.49	2.31	0.88
good-natured	5.86	1.39	2.12	0.78
goodhearted	5.96	1.40	2.23	0.88
gracious	5.60	1.27	2.02	0.89
harmony	6.40	1.08	2.02	0.91
helpful	6.12	0.86	2.06	0.80
honest	5.51	1.39	2.10	0.90
honesty	5.65	1.54	2.15	0.83
hospitable	5.94	1.44	1.65	1.09
humble	5.47	1.20	1.88	1.02
humility	5.31	1.59	1.61	1.35
influence	5.18	1.65	1.04	1.24
interdependent	6.10	1.42	0.61	1.17
interpersonal	6.27	1.28	1.21	1.29
just	5.53	1.52	2.15	0.87
kind	5.98	1.41	2.13	0.96
love	6.15	1.24	2.45	0.87
loyal	6.00	1.38	2.10	0.86
loyalty	5.96	1.55	1.96	0.95
moral	5.63	1.42	1.83	1.17
nice	5.72	1.39	1.86	0.98
nurturance	5.67	1.37	1.40	1.01
nurturing	6.21	1.20	1.77	0.98
obliging	5.28	1.46	0.09	1.28
open-minded	5.40	1.58	1.96	0.87
optimistic	5.21	1.73	2.08	0.94
patient	5.53	1.40	1.75	0.86
pleasant	5.66	1.32	1.92	0.96
polite	5.77	1.34	1.79	1.16
politeness	5.71	1.30	1.74	1.07
popular	5.63	1.30	1.16	1.26
reasonable	5.38	1.48	1.91	0.86
reliable	5.72	1.37	2.20	0.91
respectful	5.73	1.45	2.14	1.17
self-sacrificing	6.08	1.27	0.59	1.67
selfless	6.02	1.27	1.65	1.38
sensitivity	5.38	1.53	0.74	1.39
sincere	5.61	1.41	1.96	1.21
sociable	6.17	1.00	1.59	0.91

social	6.27	1.13	1.42	1.11
supportive	6.41	1.08	2.17	0.83
sympathetic	6.00	1.44	1.73	1.03
talkative	5.13	1.62	0.14	0.94
team-player	6.69	0.65	1.94	1.01
thoughtful	5.76	1.44	2.10	1.04
tolerant	5.76	1.30	1.63	1.04
trust	6.32	0.86	2.27	0.79
trusting	6.06	0.98	1.48	1.13
trustworthy	5.77	1.45	2.29	1.09
truthful	5.58	1.40	1.87	1.23
understanding	6.13	1.13	1.98	0.89
universalism	5.30	1.74	0.49	1.21
warm	6.02	1.33	1.88	0.94
warmth	6.04	1.15	2.04	0.87
welcoming	6.32	1.19	1.92	0.86
well-mannered	5.71	1.44	1.76	1.18
wise	5.21	1.70	2.10	0.96

Table S4c: Measure of inter-rater agreement for agency and communality dictionaries.

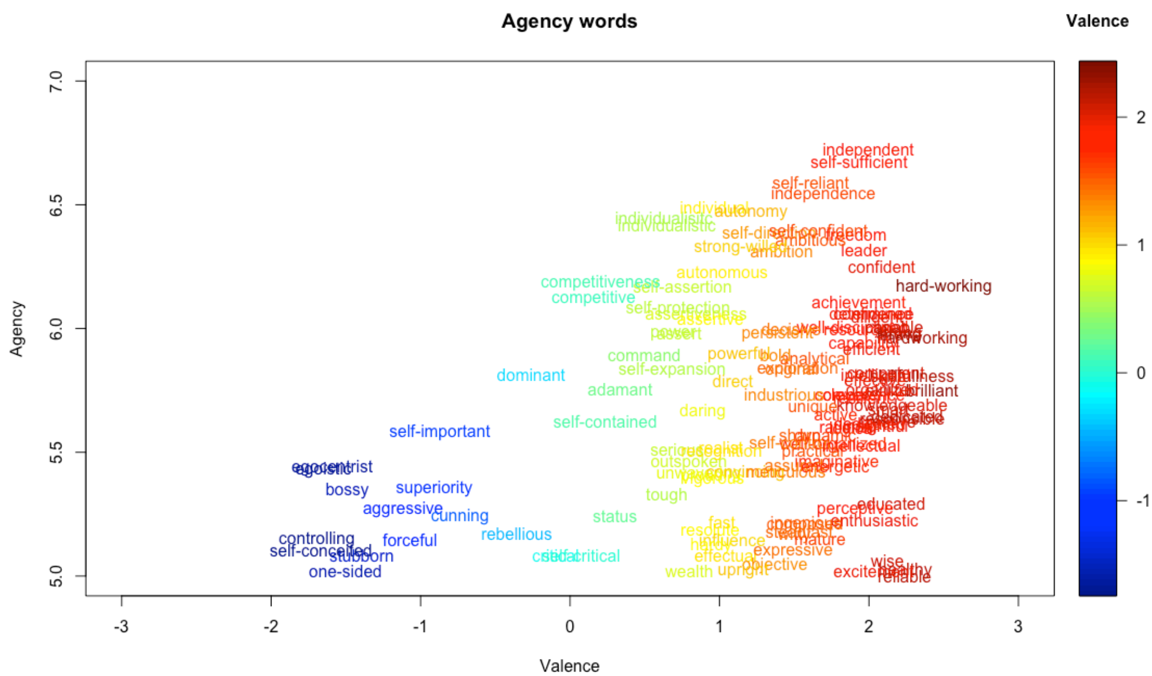
	Inter-rater agreement			
Word	Agency	Communality	Valence	Dictionary
achievement	0.617	0.011	0.738	agency
active	0.557	0.362	0.771	agency
adamant	0.526	0.291	0.586	agency
aggressive	0.237	0.615	0.475	agency
ambition	0.722	0.258	0.610	agency
ambitious	0.843	0.355	0.735	agency
analytical	0.727	0.288	0.657	agency
assert	0.589	0.310	0.620	agency
assertive	0.544	0.230	0.540	agency
assertiveness	0.570	0.284	0.497	agency
assured	0.426	0.442	0.722	agency
autonomous	0.583	0.200	0.555	agency
autonomy	0.749	0.267	0.653	agency
bold	0.540	0.206	0.665	agency
bossy	0.088	0.500	0.559	agency
brave	0.654	0.191	0.701	agency
brilliant	0.569	0.170	0.835	agency
capability	0.570	0.250	0.693	agency
capable	0.580	0.247	0.804	agency
clever	0.461	0.281	0.685	agency
command	0.561	0.245	0.572	agency
competence	0.556	0.326	0.733	agency
competent	0.535	0.544	0.767	agency
competitive	0.514	0.467	0.591	agency
competitiveness	0.535	0.266	0.483	agency
confidence	0.593	0.271	0.734	agency
confident	0.644	0.117	0.761	agency
convincing	0.449	0.164	0.712	agency
creative	0.493	0.423	0.746	agency
cunning	0.209	0.533	0.435	agency
daring	0.454	0.458	0.677	agency
decisive	0.635	0.337	0.733	agency

determined	0.549	0.249	0.783	agency
diligent	0.794	0.439	0.718	agency
direct	0.479	0.471	0.623	agency
dominant	0.397	0.378	0.581	agency
dynamic	0.661	0.462	0.784	agency
educated	0.373	0.405	0.795	agency
effective	0.511	0.257	0.792	agency
efficient	0.604	0.179	0.797	agency
egocentrist	0.246	0.620	0.559	agency
egoistic	0.250	0.539	0.583	agency
energetic	0.437	0.278	0.705	agency
exploration	0.642	0.314	0.645	agency
fast	0.201	0.205	0.707	agency
freedom	0.820	0.235	0.685	agency
hardworking	0.500	0.399	0.838	agency
imaginative	0.617	0.472	0.684	agency
independence	0.734	0.069	0.653	agency
independent	0.893	0.000	0.784	agency
individual	0.582	0.393	0.614	agency
individualistic	0.640	0.301	0.619	agency
industrious	0.627	0.212	0.501	agency
ingenious	0.446	0.407	0.556	agency
insightful	0.630	0.321	0.772	agency
intellectual	0.530	0.266	0.668	agency
intelligent	0.450	0.103	0.617	agency
knowledgeable	0.658	0.504	0.752	agency
leader	0.771	0.180	0.708	agency
logical	0.428	0.348	0.724	agency
meticulous	0.426	0.415	0.639	agency
organized	0.590	0.405	0.803	agency
original	0.603	0.256	0.645	agency
outspoken	0.458	0.207	0.658	agency
perceptive	0.364	0.195	0.698	agency
persistent	0.724	0.264	0.728	agency
power	0.682	0.309	0.514	agency
powerful	0.591	0.333	0.591	agency
practical	0.479	0.270	0.736	agency
proud	0.341	0.256	0.585	agency
rational	0.630	0.587	0.759	agency
realist	0.362	0.360	0.739	agency
recognition	0.500	0.320	0.678	agency
resilient	0.514	0.405	0.730	agency
resourceful	0.635	0.226	0.783	agency
self-assertion	0.660	0.152	0.477	agency
self-confident	0.801	0.160	0.634	agency
self-contained	0.483	0.297	0.507	agency
self-control	0.298	0.386	0.629	agency
self-direction	0.773	0.344	0.652	agency
self-expansion	0.443	0.170	0.564	agency
self-important	0.268	0.213	0.412	agency
self-protection	0.627	0.282	0.494	agency
self-reliant	0.855	0.295	0.753	agency
self-sufficient	0.899	0.056	0.782	agency
serious	0.610	0.483	0.651	agency
sharp	0.501	0.294	0.654	agency
skillful	0.516	0.276	0.763	agency
skillfulness	0.602	0.326	0.802	agency
smart	0.579	0.310	0.657	agency

status	0.515	0.314	0.545	agency
strong	0.495	0.181	0.794	agency
strong-willed	0.683	0.232	0.512	agency
superiority	0.271	0.331	0.511	agency
tough	0.496	0.243	0.481	agency
unique	0.567	0.130	0.720	agency
unwavering	0.560	0.374	0.593	agency
vigorous	0.250	0.335	0.621	agency
well-disciplined	0.585	0.402	0.743	agency
well-organized	0.567	0.443	0.583	agency
accepting	0.212	0.575	0.721	communality
affectionate	0.113	0.458	0.672	communality
agreeable	0.255	0.426	0.688	communality
altruism	0.184	0.576	0.460	communality
altruistic	0.054	0.442	0.502	communality
attachment	0.395	0.359	0.661	communality
belonging	0.417	0.624	0.728	communality
benevolence	0.200	0.212	0.553	communality
care-taking	0.090	0.752	0.708	communality
caring	0.224	0.559	0.772	communality
cheerful	0.000	0.375	0.710	communality
civility	0.369	0.800	0.798	communality
closeness	0.248	0.502	0.606	communality
communal	0.243	0.621	0.548	communality
communicative	0.087	0.554	0.782	communality
compassion	0.182	0.660	0.802	communality
compassionate	0.206	0.636	0.698	communality
compromising	0.229	0.354	0.495	communality
connected	0.172	0.801	0.741	communality
connections	0.286	0.624	0.772	communality
conscientious	0.256	0.672	0.706	communality
considerate	0.111	0.833	0.729	communality
consideration	0.344	0.734	0.760	communality
cooperation	0.099	0.887	0.794	communality
cooperative	0.153	0.835	0.838	communality
dependable	0.218	0.608	0.666	communality
dependency	0.335	0.569	0.447	communality
dependent	0.081	0.439	0.397	communality
duty	0.090	0.349	0.649	communality
easygoing	0.096	0.447	0.658	communality
enthusiastic	0.278	0.410	0.814	communality
equality	0.192	0.674	0.534	communality
fair	0.239	0.574	0.749	communality
fair-minded	0.246	0.352	0.528	communality
faithful	0.000	0.554	0.713	communality
family-oriented	0.111	0.658	0.684	communality
flexible	0.284	0.386	0.774	communality
forgiveness	0.106	0.575	0.758	communality
forgiving	0.256	0.663	0.803	communality
friendliness	0.313	0.588	0.742	communality
friendly	0.239	0.673	0.823	communality
generous	0.000	0.521	0.745	communality
gentle	0.115	0.556	0.766	communality
good	0.191	0.447	0.807	communality
good-natured	0.180	0.520	0.848	communality
goodhearted	0.094	0.511	0.806	communality
gracious	0.192	0.599	0.803	communality
harmony	0.083	0.710	0.795	communality

helpful	0.395	0.816	0.839	communality
honest	0.194	0.520	0.800	communality
honesty	0.111	0.410	0.826	communality
hospitable	0.298	0.485	0.703	communality
humble	0.065	0.643	0.738	communality
humility	0.011	0.371	0.544	communality
influence	0.213	0.316	0.617	communality
interdependent	0.021	0.497	0.658	communality
interpersonal	0.121	0.588	0.586	communality
just	0.128	0.426	0.809	communality
kind	0.245	0.505	0.770	communality
love	0.059	0.617	0.812	communality
loyal	0.004	0.521	0.817	communality
loyalty	0.000	0.396	0.776	communality
moral	0.229	0.493	0.656	communality
nice	0.167	0.520	0.760	communality
nurturance	0.243	0.533	0.747	communality
nurturing	0.174	0.642	0.759	communality
obliging	0.399	0.471	0.589	communality
open-minded	0.107	0.375	0.813	communality
optimistic	0.191	0.251	0.778	communality
patient	0.235	0.512	0.814	communality
pleasant	0.176	0.562	0.768	communality
polite	0.204	0.551	0.664	communality
politeness	0.247	0.575	0.712	communality
popular	0.309	0.578	0.601	communality
reasonable	0.284	0.451	0.817	communality
reliable	0.333	0.530	0.792	communality
respectful	0.405	0.471	0.656	communality
self-sacrificing	0.050	0.598	0.303	communality
selfless	0.000	0.599	0.527	communality
sensitivity	0.181	0.418	0.517	communality
sincere	0.199	0.502	0.633	communality
sociable	0.372	0.752	0.793	communality
social	0.262	0.679	0.693	communality
supportive	0.000	0.709	0.826	communality
sympathetic	0.213	0.478	0.737	communality
talkative	0.294	0.344	0.781	communality
team-player	0.010	0.894	0.746	communality
thoughtful	0.080	0.484	0.731	communality
tolerant	0.160	0.578	0.727	communality
trust	0.000	0.814	0.843	communality
trusting	0.097	0.762	0.681	communality
trustworthy	0.000	0.476	0.703	communality
truthful	0.230	0.512	0.624	communality
understanding	0.153	0.678	0.803	communality
universalism	0.146	0.240	0.632	communality
warm	0.025	0.560	0.781	communality
warmth	0.210	0.667	0.809	communality
welcoming	0.209	0.649	0.814	communality
well-mannered	0.338	0.479	0.654	communality
wise	0.389	0.277	0.768	communality

Figure S1: Example words plotted in agency / valence space.



Gender dictionaries. The process for recovering our gender dictionary is described in full in the Methods section of the main manuscript. Importantly, the procedure deviates slightly from our pre-registered analyses. We had initially intended to use ratings of the organizational relevance of these words collected from Amazon's Mechanical Turk to refine our gender dictionaries. However, we realized after our pre-registration – but prior to analyzing our data – that in order to isolate the association between our dictionaries and agency that could be uniquely attributed to gender, our gender dictionaries had to be 'paired'. In other words, by only including words for which perfect opposite gender analogues exist (e.g., 'he' and 'she'), we can conclude that the difference between these words' (e.g., 'he' and 'she') association with another word (e.g., 'intelligent') is due to gender. Accordingly, we refined our gender dictionaries to a subset of words from LIWC 2015's gender dictionaries (25), that were both organizationally relevant and had an opposite-gender analogue

To provide some more intuition around this decision, imagine the word "housewife", which represents one of the categories in the female dictionary. While the word "housewife" is associated with the concept of being female, it is also associated with a wide range of other topics that go beyond immediate gender associations. Capturing variance in these other latent dimensions would add large amounts of noise to the study of the relationship between gender and agency. This facilitated the analysis of **Fig. 3** – where we compare changes in the relationship between female and male dictionaries and agency words. Because these two gender dictionaries (female and male) corresponded to each other on all dimensions except gender, we were able to conclude that our effect was stronger (the interaction term) for female than male words as a result of the gendering of language.

For the 20 paired words themselves, please see below.

> male_paired_all=['mr','his','him','he','men','man','male','himself','males','masculine','masculinity','son','sons','fathers','father','paternal','brother','brothers','boy','boys']
>
> female_paired_all=['mrs','hers','her','she','women','woman','female','herself','females','feminine','femininity','daughter','daughters','mothers','mother','maternal','sister','sisters','girl','girls']

We discuss the robustness of our results to different design choices concerning the gender dictionaries later in this supplement.

Nature of text data and word use

In order to provide some more context for the nature of our text documents, and how the gendered referents are used within those documents, we provide some information regarding the text data in the following section. First, we provide excerpts showing uses of each of the female words that made up our female gender dictionary (Table S5). Second, we provide a table displaying the relative frequency of use of each of the female and agency words in the updating corpora (Table S6). Finally, we provide some visualizations (Figure S2, Figure S3) to aid in interpreting this data, breaking down the frequency of word use by document type (10-K, DEF 14A, transcripts of investor calls). This aims to help convey the different nature of word use across the documents in our corpus.

Table S5: Example context words for each of the female gender words.

Female Gender Word	Use example
Mrs	mrs park was a great global citizen but also a great icon of this country
Mrs	mrs taylors supervisory experience in financial management roles makes her an effective member of the audit committee
Hers	no transfer or assignment of employees rights an employees rights under the plan are his or hers alone
Her	denise and her team have such a strong understanding of that customer theyre not perfect they dont get it right every time
Her	thirty years of experience in investment banking extensive experience in mergers and acquisitions together with her competence in critical financial analysis and successful record in a variety of business dealings
She	she went on to serve as dean of the graduate school of arts and sciences

She	she gets inside the company and understands better how we do things shes incredibly excited she was very successful
Women	the board is committed to actively seeking highly qualified women and individuals from minority groups to include in the pool from which board nominees are selected
Women	we are committed to creating an environment in which women feel empowered by both their inner and outer beauty
Woman	as an asian american woman ms gin adds to the diversity of the boards and enhances exelons diversity initiatives and community outreach
Woman	we believe this event is a way to empower woman to strengthen the emotional command and in our brand while giving to an important cause
Female	we exceeded our target that of new employees who joined our top three career bands were female and or minorities on growth relative to our peer group
Female	ms grecos role as a female executive has brought diversity to pecos board and she has contributed to pecos diversity initiatives
Herself	heather has immersed herself in every key area of our business and distinguished herself as the leader within the generics industry she has unmatched knowledge of mylan and the industry
Herself	in a welcoming and approachable environment our customer can immerse herself in our extensive product selection indulge herself in our hair or skin treatments or discover new and exciting products in an interactive setting
Females	the companys percentage of females declined overall largely due to workforce reductions in the phone business while it increased slightly for females in technical fields
Females	and you can see some of the targets there that were placing the ads in were targeting females from to and in addition to that as you can well imagine weve got a heavy internet presence
Feminine	the company gender singular and plural all pronouns and any variations thereof will be deemed to refer to the masculine feminine or neuter as the identity of the person or persons may require
Feminine	feminine care products business unit various positions in general and brand management both in the united states and internationally
Daughter	i talked about video editing with my daughter earlier sony vegas pro is an enthusiast professional video editing application it does some very cool things
Daughter	the daughter of mr rj leblanc senior vice president ibm cloud are also employed by ibm in nonexecutive positions

Daughters	i think as parent you should do that i have two daughters as well and its important that we understand in an online world who our family is interacting with
Daughters	ronald s lauders daughters will succeed to their fathers rights if he should cease to be a director by reason of his death
Mothers	ibm supports protects and empowers its employees expansive family leave provides new birth mothers with up to weeks of paid leave and weeks for fathers partners
Mothers	migraine both episodic and chronic which is a crippling disease resulting in mothers not being able to be mothers or employees not being able to do their work
Mother	our approach continues to earn top recognition from leading publications and organizations including diversity journal working mother equal opportunity minority engineer and diversity mba magazines
Mother	i am a mother and a grandmother i take this role very seriously as a guardian and a protector of the future
Maternal	demonstrating the impact incentive enhanced and increase participation in a maternity management program and improving maternal and infant health
Maternal	roughly of pregnancies today are afflicted by preeclampsia and nearly of maternal deaths and of premature deaths are caused by preeclampsia
Sister	we made a discretionary decision to defer the launch of our new mondeo product which is the sister product of the fusion
Sister	is to have the directors each one be assigned a product to be a big brother for or sister so that they focus on that one they can drive that product
Sisters	successive preference a surviving spouse if any b surviving children equally c surviving parents equally d surviving brothers and sisters equally
Sisters	i well remember my sisters used to push me to the front in front of my parents to ask for something
Girl	in that positioning and its really targeting to that yearold collegiate girl so then it can also go down in age as well
Girl	we launched new creative for our television and radio ads centered on our joy to the girl campaign which was integrated across all touch points
Girls	i think what delivery promises in these very early test markets the guys and girls are actually running this on a daily basis are seeing that benefit were encouraged by that
Girls	oracle will offer more than direct educational events and support conferences summer computing camps and codefests for girls with the aim of inspiring them to explore and pursue opportunities in stem fields

Note. The pre-registered female word "femininity" occurred 0 times, so is not included. The word "hers" occurred once, so only one example is provided. For all other words, we provide two excerpts.

Table S6: The total count and relative frequency of each of our focal words in the updating corpora (39 focal organizations across 2 periods).

Word	Dictionary	Count	Relative Frequency
effective	Agency	41088	0.05%
independent	Agency	36597	0.07%
power	Agency	34485	0.04%
determined	Agency	33337	0.05%
strong	Agency	30985	0.05%
individual	Agency	20352	0.03%
competitive	Agency	17806	0.03%
direct	Agency	14900	0.02%
independence	Agency	8923	0.02%
active	Agency	8436	0.01%
recognition	Agency	7830	0.01%
original	Agency	7282	0.01%
status	Agency	6983	0.01%
achievement	Agency	6118	0.01%
intellectual	Agency	6032	0.01%
unique	Agency	4962	0.01%
confidence	Agency	3758	0.01%
confident	Agency	3693	0.01%
efficient	Agency	3388	0.01%
capability	Agency	3382	0.01%
leader	Agency	2824	0.00%
smart	Agency	2378	0.00%
tough	Agency	2067	0.00%
aggressive	Agency	1962	0.00%
dynamic	Agency	1819	0.00%
fast	Agency	1700	0.00%
analytical	Agency	1501	0.00%
proud	Agency	1314	0.00%
competitiveness	Agency	1255	0.00%
organized	Agency	1253	0.00%
practical	Agency	1095	0.00%
competent	Agency	1076	0.00%
autonomous	Agency	799	0.00%
capable	Agency	778	0.00%
assured	Agency	682	8.00E-06
intelligent	Agency	668	9.00E-06
serious	Agency	659	9.00E-06
assert	Agency	504	5.00E-06
creative	Agency	497	8.00E-06
command	Agency	451	7.00E-06
sharp	Agency	440	7.00E-06
rational	Agency	421	7.00E-06
freedom	Agency	377	6.00E-06
exploration	Agency	295	3.00E-06
bold	Agency	271	4.00E-06

dominant	Agency	259	3.00E-06
diligent	Agency	230	3.00E-06
resilient	Agency	208	3.00E-06
persistent	Agency	144	2.00E-06
logical	Agency	143	3.00E-06
knowledgeable	Agency	137	2.00E-06
ambitious	Agency	131	2.00E-06
superiority	Agency	127	2.00E-06
competence	Agency	119	2.00E-06
decisive	Agency	100	2.00E-06
autonomy	Agency	95	1.00E-06
unwavering	Agency	89	2.00E-06
vigorous	Agency	75	1.00E-06
convincing	Agency	74	1.00E-06
ambition	Agency	73	1.00E-06
brilliant	Agency	72	1.00E-06
energetic	Agency	65	1.00E-06
insightful	Agency	35	1.00E-06
educated	Agency	25	0%
hardworking	Agency	24	1.00E-06
resourceful	Agency	18	0%
clever	Agency	16	0%
outspoken	Agency	14	0%
adamant	Agency	9	0%
perceptive	Agency	8	0%
brave	Agency	6	0%
assertive	Agency	5	0%
meticulous	Agency	3	0%
realist	Agency	3	0%
imaginative	Agency	2	0%
individualistic	Agency	2	0%
bossy	Agency	1	0%
ingenious	Agency	1	0%
skillful	Agency	1	0%
assertiveness	Agency	0	0%
cunning	Agency	0	0%
daring	Agency	0	0%
egocentrist	Agency	0	0%
egoistic	Agency	0	0%
industrious	Agency	0	0%
powerful	Agency	0	0%
self-assertion	Agency	0	0%
self-confident	Agency	0	0%
self-contained	Agency	0	0%
self-control	Agency	0	0%
self-direction	Agency	0	0%
self-expansion	Agency	0	0%
self-important	Agency	0	0%
self-protection	Agency	0	0%
self-reliant	Agency	0	0%
self-sufficient	Agency	0	0%
skillfullness	Agency	0	0%
strong-willed	Agency	0	0%
well-disciplined	Agency	0	0%
well-organized	Agency	0	0%
fair	Communality	71275	0.08%
just	Communality	65625	0.12%
good	Communality	42747	0.08%

kind	Communality	21861	0.04%
reasonable	Communality	15938	0.02%
trust	Communality	15722	0.02%
consideration	Communality	12829	0.02%
social	Communality	4834	0.01%
understanding	Communality	4759	0.01%
dependent	Communality	3866	0.00%
patient	Communality	2497	0.00%
helpful	Communality	2484	0.00%
nice	Communality	2253	0.00%
duty	Communality	2209	0.00%
influence	Communality	1843	0.00%
loyalty	Communality	1746	0.00%
reliable	Communality	1549	0.00%
accepting	Communality	1449	0.00%
sensitivity	Communality	1340	0.00%
optimistic	Communality	1312	0.00%
connected	Communality	1300	0.00%
love	Communality	1155	0.00%
flexible	Communality	986	0.00%
cooperation	Communality	944	0.00%
attachment	Communality	614	7.00E-06
popular	Communality	497	8.00E-06
thoughtful	Communality	447	8.00E-06
connections	Communality	414	5.00E-06
honest	Communality	367	7.00E-06
supportive	Communality	336	5.00E-06
cooperative	Communality	315	3.00E-06
moral	Communality	242	4.00E-06
wise	Communality	215	4.00E-06
enthusiastic	Communality	200	4.00E-06
friendly	Communality	195	4.00E-06
dependable	Communality	183	3.00E-06
warm	Communality	175	3.00E-06
loyal	Communality	165	3.00E-06
honesty	Communality	125	2.00E-06
truthful	Communality	118	1.00E-06
interdependent	Communality	93	1.00E-06
belonging	Communality	88	1.00E-06
forgiveness	Communality	81	1.00E-06
welcoming	Communality	81	1.00E-06
caring	Communality	74	1.00E-06
equality	Communality	74	1.00E-06
pleasant	Communality	66	1.00E-06
compromising	Communality	58	1.00E-06
respectful	Communality	58	1.00E-06
agreeable	Communality	57	1.00E-06
dependency	Communality	57	1.00E-06
humble	Communality	40	1.00E-06
generous	Communality	38	1.00E-06
compassionate	Communality	27	0%
humility	Communality	25	0%
interpersonal	Communality	25	0%
trustworthy	Communality	21	0%
sincere	Communality	20	0%
nurturing	Communality	14	0%
conscientious	Communality	13	0%
trusting	Communality	13	0%

faithful	Communality	12	0%
friendliness	Communality	9	0%
gentle	Communality	9	0%
tolerant	Communality	9	0%
closeness	Communality	8	0%
compassion	Communality	6	0%
gracious	Communality	5	0%
harmony	Communality	5	0%
sympathetic	Communality	5	0%
forgiving	Communality	4	0%
cheerful	Communality	2	0%
communicative	Communality	2	0%
considerate	Communality	2	0%
hospitable	Communality	2	0%
altruistic	Communality	1	0%
communal	Communality	1	0%
politeness	Communality	1	0%
warmth	Communality	1	0%
affectionate	Communality	0	0%
altruism	Communality	0	0%
benevolence	Communality	0	0%
care-taking	Communality	0	0%
civility	Communality	0	0%
easygoing	Communality	0	0%
fair-minded	Communality	0	0%
family-oriented	Communality	0	0%
good-natured	Communality	0	0%
goodhearted	Communality	0	0%
nurturance	Communality	0	0%
obliging	Communality	0	0%
open-minded	Communality	0	0%
polite	Communality	0	0%
self-sacrificing	Communality	0	0%
selfless	Communality	0	0%
sociable	Communality	0	0%
talkative	Communality	0	0%
team-player	Communality	0	0%
universalism	Communality	0	0%
well-mannered	Communality	0	0%
her	Female	17638	0.03%
she	Female	8256	0.01%
mrs	Female	841	0.00%
women	Female	636	0.00%
herself	Female	241	4.00E-06
female	Female	176	3.00E-06
mothers	Female	139	3.00E-06
feminine	Female	132	2.00E-06
girl	Female	101	2.00E-06
daughters	Female	94	2.00E-06
woman	Female	84	1.00E-06
girls	Female	68	1.00E-06
sister	Female	68	1.00E-06
sisters	Female	68	1.00E-06
females	Female	63	1.00E-06
mother	Female	54	1.00E-06
daughter	Female	47	1.00E-06
maternal	Female	2	0%
hers	Female	1	0%

femininity	Female	0	0%
mr	Male	67204	0.13%
his	Male	43948	0.07%
he	Male	25642	0.05%
him	Male	4218	0.01%
men	Male	669	5.00E-06
himself	Male	414	7.00E-06
brothers	Male	221	3.00E-06
sons	Male	218	3.00E-06
man	Male	210	3.00E-06
son	Male	180	3.00E-06
brother	Male	147	2.00E-06
masculine	Male	145	2.00E-06
male	Male	99	1.00E-06
boy	Male	81	1.00E-06
males	Male	63	1.00E-06
father	Male	49	1.00E-06
boys	Male	47	1.00E-06
fathers	Male	27	1.00E-06
masculinity	Male	0	0%
paternal	Male	0	0%

Note. The counts refer to the corpus of documents used to update our pre-trained model to the word2vec models used in Study 1's analysis. To compute these numbers, we first took a total sum of the frequency of use of each word (the column, "Count"). To compute the "Relative Frequency" column, we first computed the relative frequency of occurrence of each word in each firm's updating corpus. This is expressed as a percentage of total words. We then averaged these frequencies across firms to compute an overall aggregate. We chose to do this to accurately reflect the across firm average, thus avoiding the data being biased by firms with larger updating corpora.

Figure S2: The most frequently occurring gender words faceted by document type.

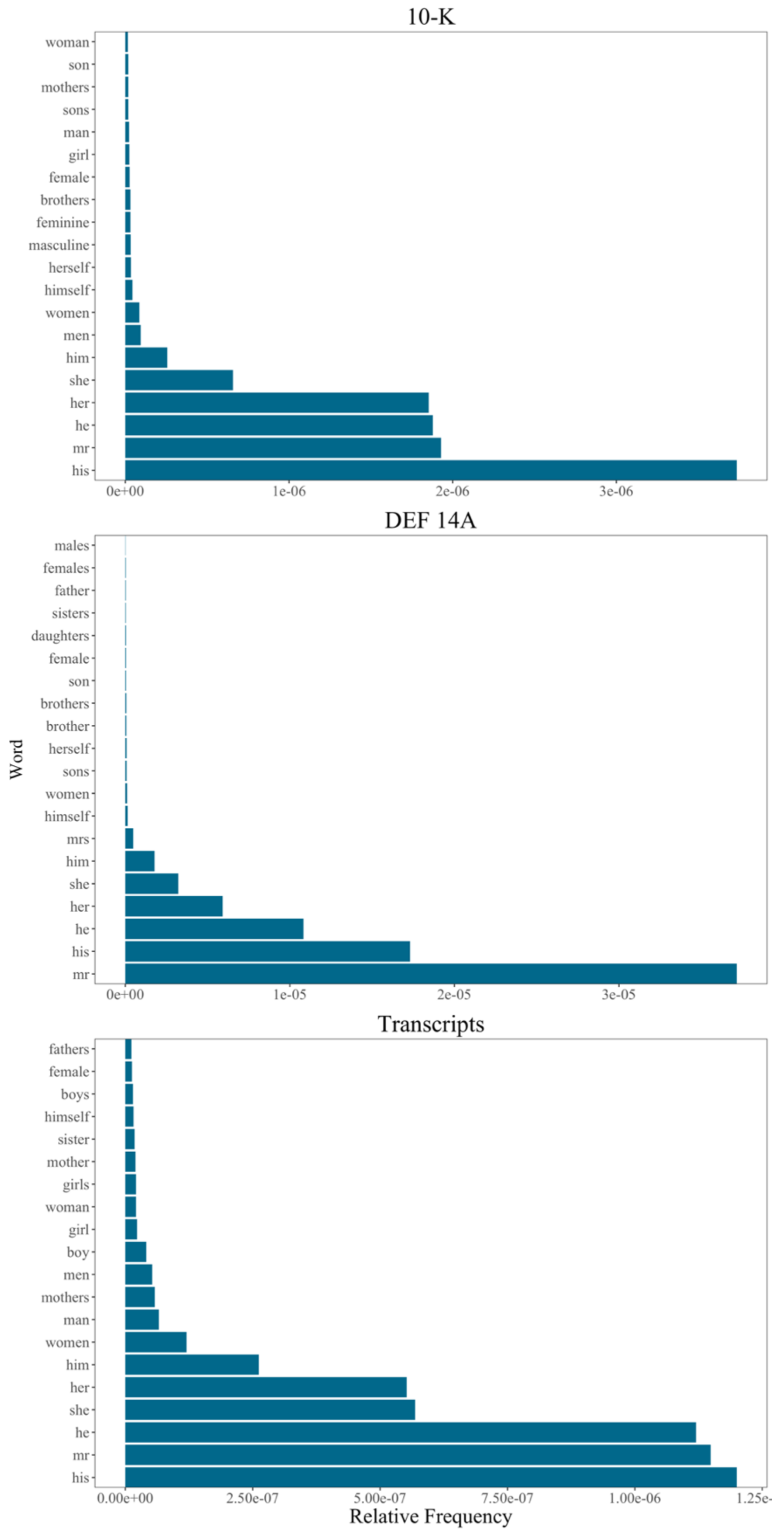
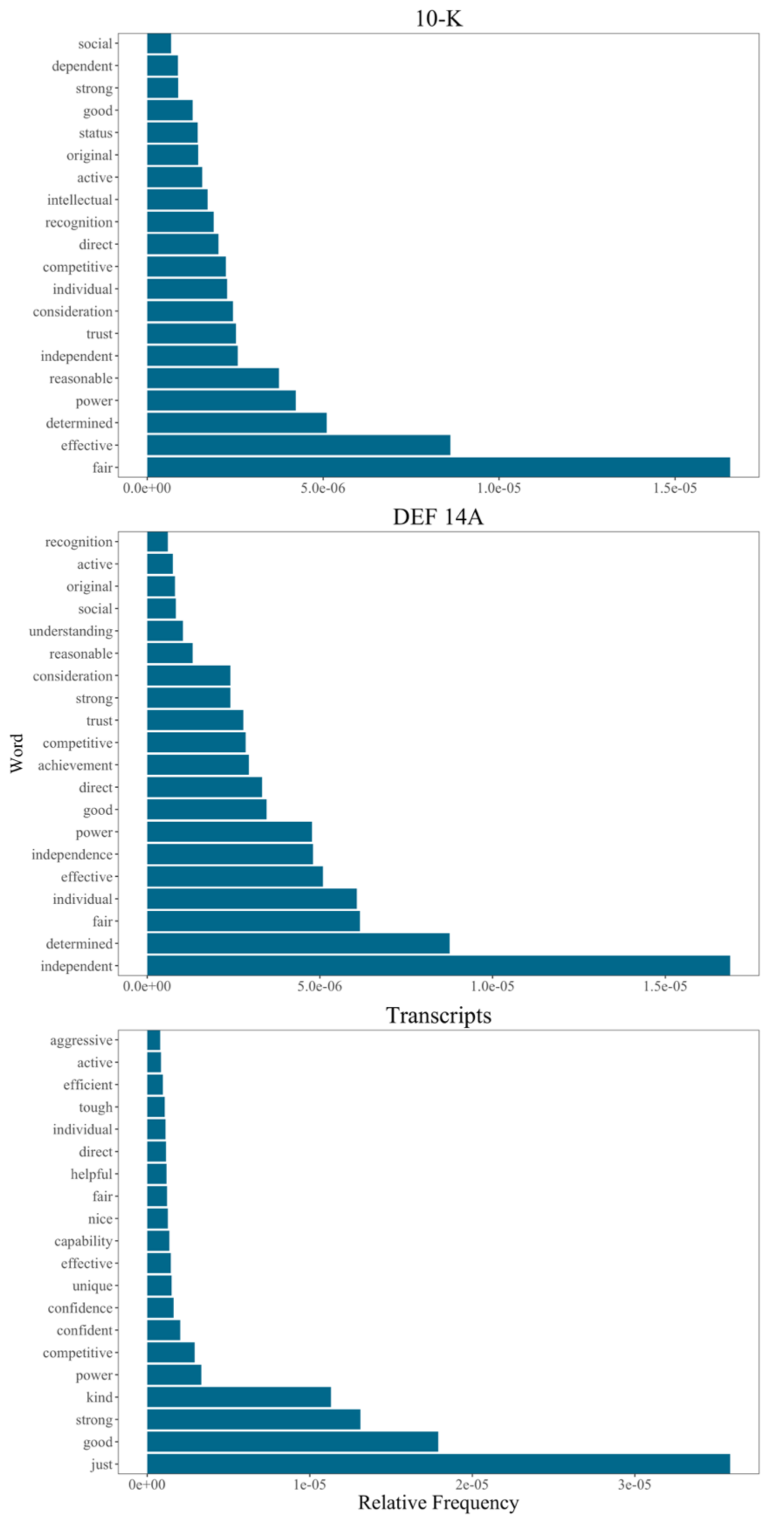


Figure S3: The most frequently occurring theory (agency, communality) words, faceted by document type.



Details of word2vec training and robustness checks

We trained word2vec models using organizational documents in order to study how the semantic relationships between words in vector space changed as a function of appointing a female CEO. In the following section, we delineate details of the procedure by which we trained these models, as well as additional analysis testing the robustness of our results to changes to this training procedure.

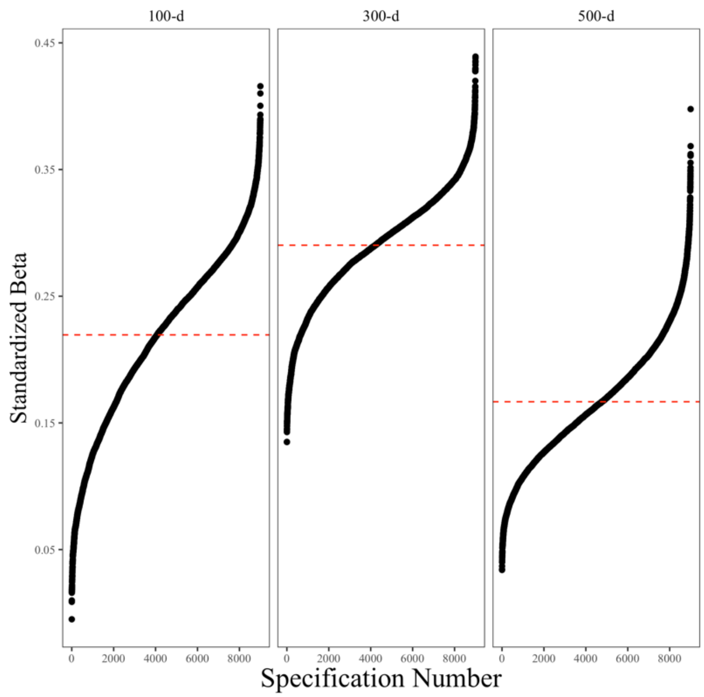
We trained our word2vec models using the 'gensim' package for Python (26). This involved the specification of several hyperparameters. First, note that we elected to estimate our models with three different sizes of word vectors (using 100, 300, and 500 dimensions). Details of how the results varied across vector size are included later in this section. We used the Continuous Bag of Words implementation of the word2vec algorithm, and implemented the negative sampling procedure. As part of the negative sampling procedure, we specified that 5 "noise words" should be drawn. We set the minimum frequency of occurrence of a word to be 2 times. For all models, we downsampled the 40 highest frequency words. The window for the bag of words used to predict focal words was 5 words either side.

We trained all models for 10 epochs using 4 workers. For the base models containing the full text sample (SEC documents and transcripts), the training times were 9.62 hours (100-d), 12.56 hours (300-d), and 14.31 hours (500-d). The updated models using these pre-trained models took a trivial amount of time to train.

Given the complexity of the training process of these models, we aimed to also provide a broad grid of specifications to identify whether our results were robust to design choices. These included varying the choice of word vector dimensionality, the choice between the hierarchical softmax and negative sampling procedure, varying the size of window used to learn the meaning of focal words, and testing different random initializations of weights between neurons. Whilst we varied the word vector size in the core results of the paper, the other changes to hyperparameters are solely included in this supplement as further robustness checks for the interested reader.

Faceted by vector size. Below we provide the main specification curve **Fig. 3** faceted by word vector size (Figure S4). The average beta estimates of the interaction effect for each word vector size were 0.219 (size 100), 0.290 (size 300), and 0.167 (500). The curvilinear pattern observed between these three datapoints may suggest that 100 dimensions is insufficient to capture the complexity in our large dataset, whereas 500 dimensions introduces significant noise from the larger number of parameters to estimate. Importantly, we find evidence of the same effect in each case. This suggests that our results are not an artefact of the design choice regarding vector size or any particular random initialization. We seek further evidence to verify this second claim in the proceeding section.

Figure S4: Specification curve plotting the standardized beta of the interaction term β3 from the regression equation 1.1 faceted by size of word vectors (100, 300, 500).



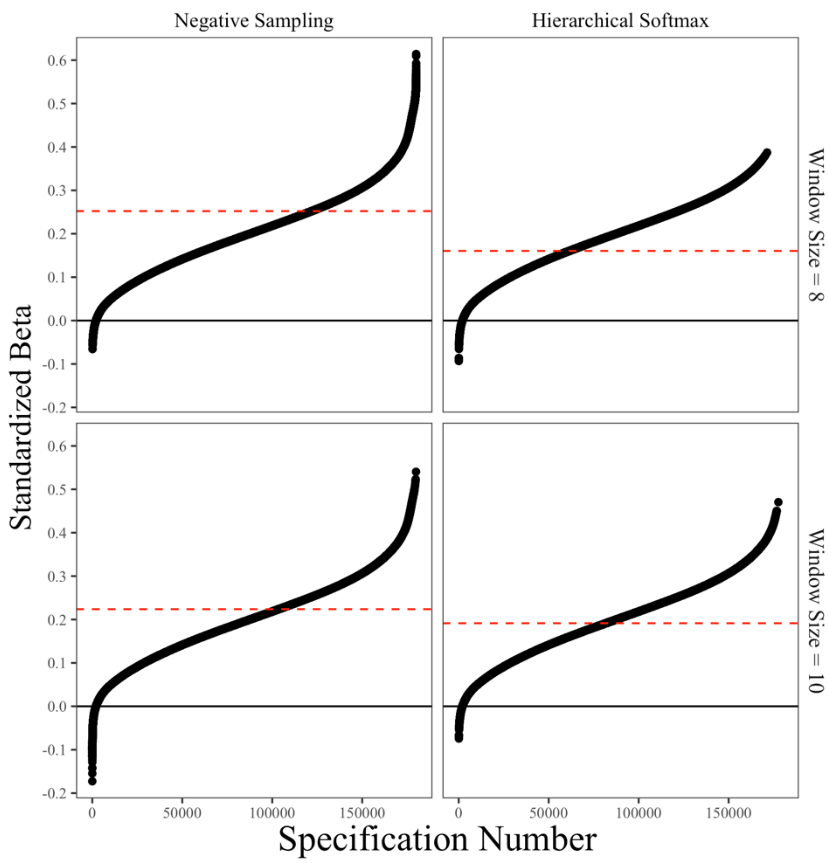
Iterating over hyperparameters and initializations

We wanted to establish that our results were not an artefact of aspects of the word2vec model training process. To do so, we conducted additional analysis to test whether the main study effect (the average standardized beta from our regression equation 1.1) varied with different design specifications. Specifically, we estimated 20 different sets of 300-dimensional word2vec models across a grid of three design choices; hierarchical softmax vs. negative sampling (2 options) X sampling window length of 8 vs. 10 (2 options) X five random initializations. We estimated all of the firm-period word embedding models using this grid of parameters, bootstrapped 1,000 samples for each specification, and computed 9 different measures of the cosine similarity between female and agency words (using the 3 different sized gender dictionaries and 3 different sized agency dictionaries) for each bootstrapped sample. Overall, this process produced 180,000 different estimates of the standardized beta coefficient. This process sought to verify that our results were not dependent on a particular random initialization, or configuration of hyperparameters. The results are included in full in Figure S5.

The average standardized beta across all specifications was $\beta = 0.207$ ($t = 844.41$ $df = 179{,}999$, $p < 0.001$). In general, estimates of the beta coefficient were somewhat higher when using negative sampling versus hierarchical softmax (average(β) = 0.238 versus average(β) = 0.176), were not affected by the window size (average(β) = 0.206 for window = 8; average(β) = 0.208 for window = 10), shrunk slightly with larger gender dictionaries (average(β) = 0.224 for female dict = 10; average(β) = 0.185 for female dict = 20), did not change with different agency dictionary sizes (average(β) ranged from 0.206 to 0.214)), and varied to some extent with different random initializations. Given the random initializations were also specific to a set of hyperparameters, the best comparison is within each of the four main specifications we tested here. The largest variance across initializations was for window size = 8 using negative sampling, where estimates of standardized beta ranged from 0.084 to 0.420. Across the 180 unique configurations of design choices, the range of the average standardized beta was from 0.024 to 0.493 (with the average beta reported in the manuscript being 0.226). Notably, none of the specifications produced an average non-positive standardized beta.

In sum, we find that the observed effect of hiring a female CEO on agency is robust across a wide grid of word2vec training specifications.

Figure S5: Grid of different specifications and the sensitivity of the standardized beta to design choices.



Further details of manuscript analysis

Pre-registration. We deviated from our pre-registered analysis plan on two points. First, we decided to use paired gender dictionaries after our pre-registration but prior to analyzing the data. The reasoning behind this is explained above in the gender dictionaries section – but can be summarized by the desire to study the differences between the female-agency and male-agency associations that could solely be explained by gender. Comparing lists of gender words that also vary on other dimensions (e.g., comparing ("she", "her") to ("man", "brother")) runs the risk of interpreting variance in different latent dimensions, rather than differences driven by gender. To this end, we focused on paired gender dictionaries as the most valid input parameter for our desired analyses.

Secondly, we elected to draw bootstrapped samples of our firms in order to present a fuller picture of variation in our analyses. Our bootstrapping procedure was the following; a) include all 11 target firms in every sample, b) for each target firm randomly select whether to include propensity match 1 twice, propensity match 1 once & propensity match 2 once (this was twice as likely to account for the two possible orders), or propensity match 2 twice, and c) run the core regression for each of our specifications using this randomly drawn sample. We elected to use this process to mitigate the arbitrariness in our classification of propensity matches into two groups (1 and 2). By bootstrapping 1,000 samples from the space of 3^{11} = 177,147 possible configurations, we recovered a balanced estimate of the effect of hiring a female CEO across a wide range of possible comparison groups. We include the results estimated without the bootstrapping procedure as a robustness check in this supplement. Our conclusions are unaffected.

Core result. The results in **Fig. 3** in the main manuscript were estimated by running the regression equations in 1.1. and 1.2 for 27 different specifications for 1,000 bootstrapped samples, for both the female and male dictionaries (listed in full earlier in this supplement). Our 27 specifications were derived from: 3 different sizes of gender dictionary (10, 15, 20), 3 different sizes of agency dictionary (50, 75, 100), and 3 different sizes of word vector dimensions (100, 300, 500). All analyses were pre-registered in February 2020. Our focal effect of interest is the interaction between the two time periods ("pre" vs. "post) and the firm condition ("target" vs. "propensity-matched") on the similarity in meaning between female and agency-related words. To reduce researcher degrees of freedom and provide a robust estimate of the effect, we took these extensive measures to estimate a wide range of β_3 coefficients when using different model specifications.

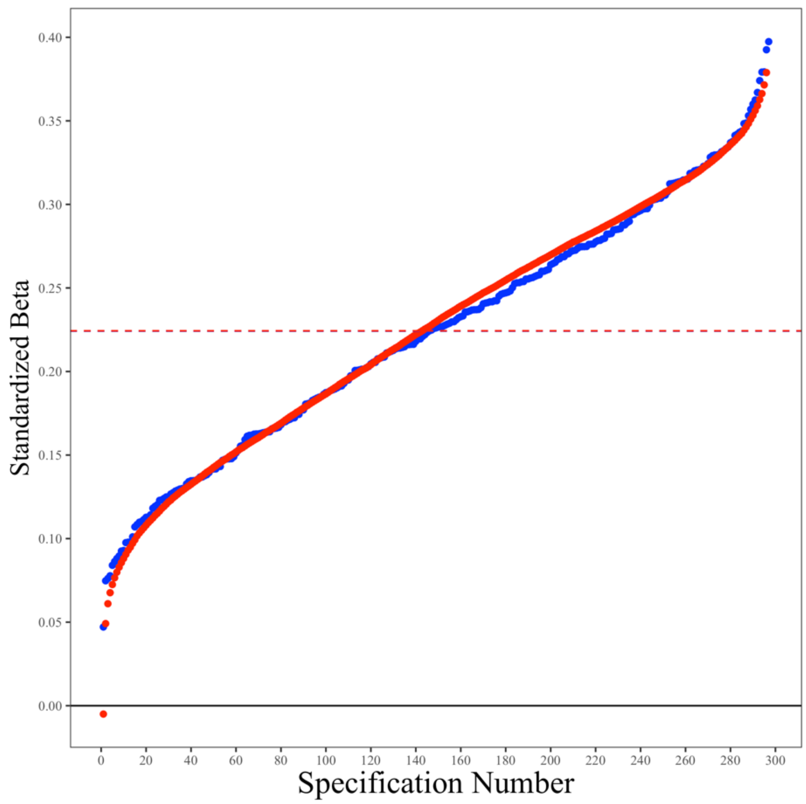
In **Fig. 2**, we plotted the average values of our measure (the association between female vectors and agency vectors) for each firm before and after the hiring date of their female CEO. These were first averaged across our 27 specifications, and then grand mean-centered across all 66 measures.

Robustness checks

In order to verify the robustness of our effect to changes in specification and analytical strategy, we implemented a wide range of robustness checks. First and foremost, all of our analyses focus on sets of specifications, rather than any one arbitrary set of parameters. For our core result, we look at the variation in the coefficient on the interaction term β_3 across 27 different configurations of agency dictionary sizes, gender dictionary sizes, and word vector sizes. Furthermore, we sought to establish our effect's generalizability and the effect of our statistical procedures on our results. This included three primary analyses: Drop one analysis within our sample of firms, estimating the standardized betas for each specification without using bootstrapping, and estimating the results using just the SEC text data.

Drop one analyses. As an unfortunate result of the lack of female representation in the S&P 500, our sample of firms is small. This opens the possibility that one firm in particular could significantly affect the results. We wanted to rule out this possibility, to ascertain that our results can facilitate conclusions about large companies that hire female CEOs more broadly, rather than any specific firm. To this end, we estimated our regression models for each of the 27 specifications using samples that dropped each of the 11 target firms (and their associated propensity matches). That is, we estimated 297 (27 specifications X 11 firms, dropped one at a time) regressions containing 60 observations each – two time periods of the 10 remaining target firms and their 20 propensity matches. The distribution of 297 betas were all positive and differed significantly from 0 (average $\beta = 0.224$, $t = 51.98$ $df = 296$, $p < 0.001$). The results of this analysis can be seen in Figure S6.

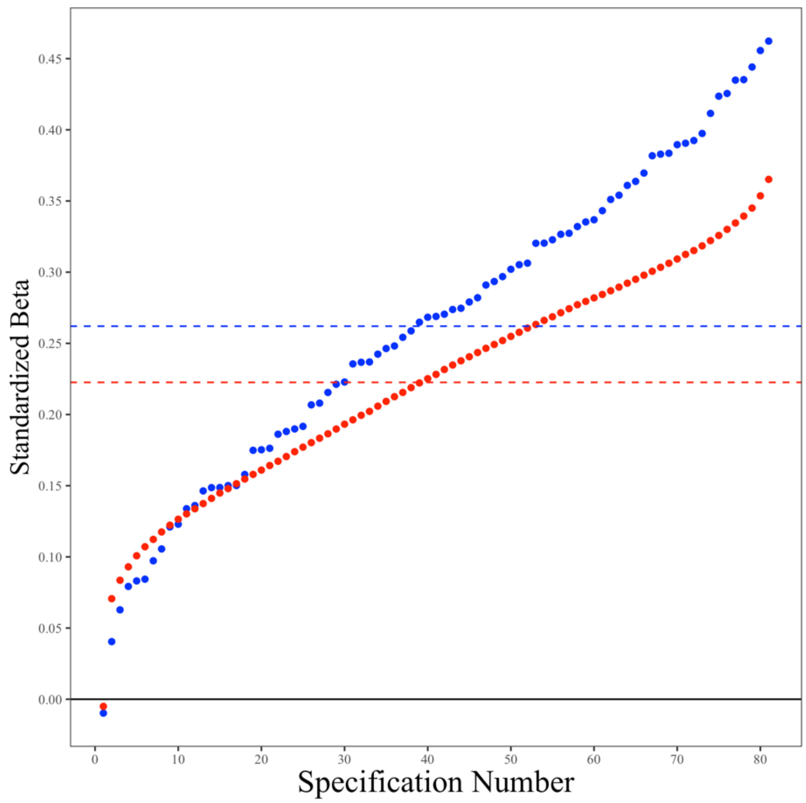
Figure S6: Specification curve plotting the standardized beta of the interaction term β3 from the regression equation 1.1 estimated for each of 27 specifications dropping 1 of the 11 target firms at a time.



Note. Red points plot the original specification curve. Blue points plot the curve dropping one organization and its propensity matches at a time.
Note. We systematically sampled points from the original specification curve to represent the curve on the same axes.

Without bootstrapping. We note our reasons for deviating from our pre-registered intention to divide our regression analyses into three categories (propensity match set 1, set 2, and both) in favor of bootstrapping, but as such it is important to show that this does not affect the conclusions we make from our data, nor the statistical significance of our effect. Hence, we include the results of these 81 regression models in Figure S7. When we use this pre-registered categorization of firms (3 possible groupings; propensity match set 1, propensity match set 2, and both), 98.8% of the betas on the interaction term were positive, differing significantly from 0 (t = 21.39, df = 80, p < 0.001). The average estimate of our effect of interest (the beta coefficient on the interaction term) is actually *larger* (average $\beta_3 = 0.262$, $SD(\beta_3) = 0.110$) using this modelling specification. However, we believe that the bootstrapped regressions represent our results with greater fidelity and therefore report the bootstrapped results in the main manuscript.

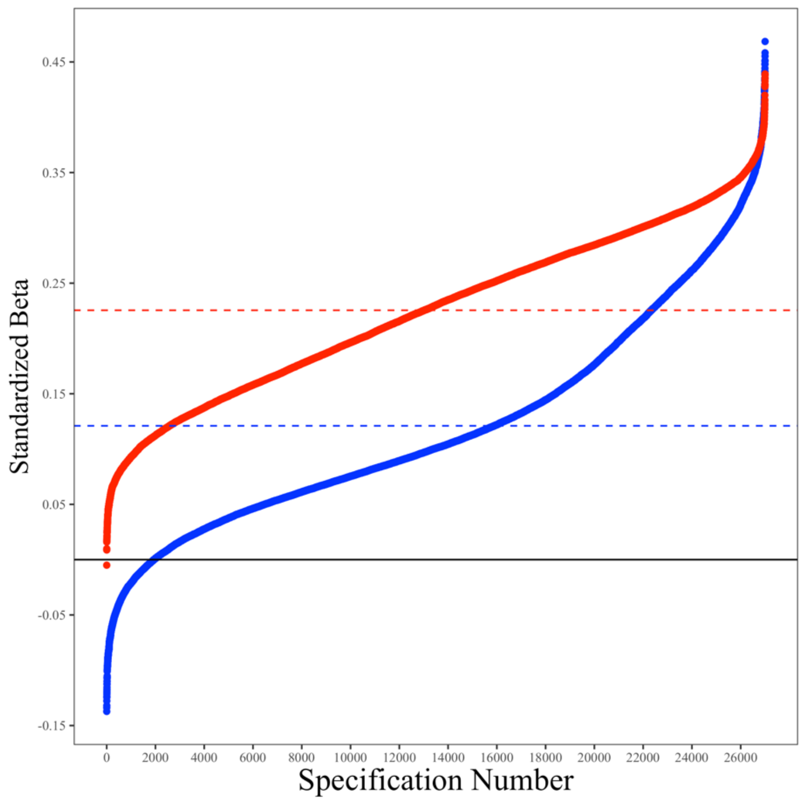
Figure S7: Specification curve plotting the standardized beta of the interaction term β3 from the regression equation 1.1 estimated for each of 27 specifications with 3 groups of propensity matches (set 1, set 2, both)



Note. Red points plot the original specification curve. Blue points plot the curve with no bootstrapping.
Note. We systematically sampled points from the original specification curve to represent the curve on the same axes.

Text source. Another possible contributing factor to our result is the nature of the training data – we include two different types of text, formal filings with the SEC, and transcripts of company investor calls. Within the SEC data, we focus on two specific types of filing, the DEF14-A filing for shareholder votes, and the 10-K form that constitutes a company's annual report. These documents differ in their nature, and thus partitioning the data based on text source is likely to affect our results. However, looking within a particular group of documents we also reduce the data available to update our word2vec models. Narrowing our dataset offers a trade-off between greater granularity and lower reliability. Consequently, we emphasize the need for caution in interpreting the results using *just* the SEC data, but take them to provide evidence of convergent validity, given we still find a positive interaction term, the beta of which differs significantly from 0 (average $\beta = 0.121$, $t = 204.38$ $df = 26{,}999$, $p < 0.001$). The resulting specification curve can be seen in Figure S8. Whilst the effect size is somewhat smaller, this is difficult to interpret due to the aforementioned issues with model training. Note that it did not make practical sense to train a model with *just* the transcripts data, as this would result in updating each of our firm-period level corpora with under 250,000 words of text, which is insufficient to expect reliable estimates.

Figure S8: Specification curve plotting the standardized beta of the interaction term β3 from the regression equation 1.1 when using only the SEC data to pre-train and update the word2vec models.



Note. Red points plot the original specification curve. Blue points plot the curve estimated with just the SEC data.

Process evidence

One possibility is that the observed increase in similarity between the meaning of women and agency is merely a product of female CEOs being described performing agentic actions as part of their role. As we have outlined in the manuscript, the event of a woman being hired as CEO means that public organizational documents are likely to refer to her and her actions as CEO. Hence, rather than companies strategically signaling the virtues of their female leader or incidentally reflecting attitude change towards women, the observed effects might simply be driven by organizations describing the actions of their female CEOs in the post-hiring period. Whilst this may be a contributing factor, there are several empirical reasons to believe that our effects are not merely driven by organizations changing from describing the actions of a male CEO to describing similar actions performed by a female CEO. We provide a summary of the relevant empirical results below, before expanding fully in the proceeding section.

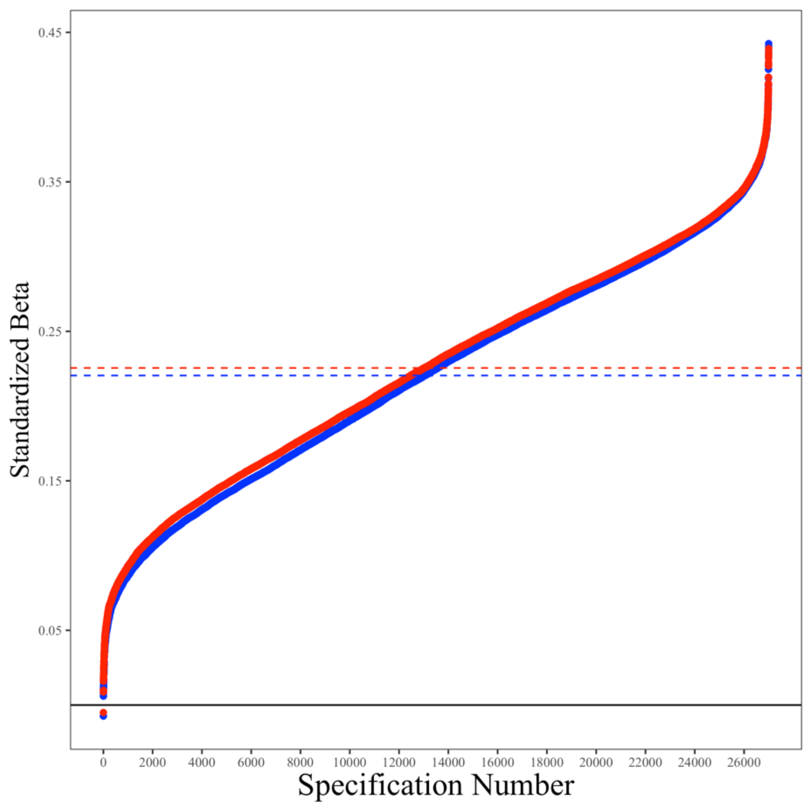
First, we do not see a negative effect on the male-agency association when female CEOs are hired (**Fig. 3**), which would be predicted if our results were purely generated by changes in the gender of the actor who is described performing the role of the CEO in our data. Moreover, adding the main effect of the relative frequency of female references in our training documents as a control variable to the regression equation 1.1 did not notably impact the focal interaction effect (average $\beta_3 = 0.220$, $SD(\beta_3) = 0.077$, $t = 470.2$, $df = 26{,}999$, $p < 0.001$, Figure S9). If the interaction effect was largely driven by organizations describing their [female] CEOs' actions without any convergence in the meaning of womanhood and agency, we would expect that organizations which used more female references would have vectors that changed more, and hence that including a variable for frequency of female references would shrink this effect.

In addition to the lack of a cost to men's agency when men are replaced as CEOs and the insignificant effect of controlling for the frequency of female references, we can gain insight into the nature of the effects by inspecting the agency words that moved the most towards the meaning of being a woman. If our results were solely explained by organizational documents describing the actions of a now female CEO, we would expect the top words to be action verbs. Contrary to this, the words which saw the largest increase in their association with the semantic meaning of women were primarily adjectives (e.g., tough, original, decisive). This provides further evidence that our results are not solely driven by description of the CEO's actions, but rather by the language used by organizations characterizing women as more agentic.

Additional analysis identified which individual words moved by the largest magnitude between the pre and post hire periods, by aligning the vector spaces of the pre and post models for each organization. If the observed similarity changes were solely driven by descriptions of CEO behaviors, we might expect to observe a larger change in the meaning of being female in target organizations, as references to female CEOs would occur in a wide range of contexts that would shift the meaning of womanhood. However, this was not the case. The female vectors did not change significantly differently in the target organizations versus the propensity matches ($p = 0.867$), and the agency vectors actually changed by less ($p < 0.001$). This evidence suggests that the effects were driven by a specific kind of meaning change, whereby leadership congruent characteristics became more similar in meaning to womanhood, rather than a general change in semantic meanings, or instability in the concept of gender. Overall, the empirical evidence is suggestive that the similarity changes were caused by more than just descriptions of CEO behavior. We provide full details of these analyses below

Female referents. One possible explanation of our results is that the effect is solely driven by a higher relative frequency of female referents in firms that hire female CEOs. That is, firms that hire female CEOs talk about their CEOs (who happen to now be female), and this leads to an increase in the association between being female and being agentic in their company text. There are at least two reasons to believe that the effect is not solely driven by this frequency change. First, the relative frequency of female referents (i.e. how many female words per word of text in the training corpora) did not differ significantly either in the period pre (t = -1.34, df = 18.783, p = 0.195) or post (t = -1.47, df = 21.956, p = 0.156) the hire of a female CEO between target and propensity matched companies. Second, when we include this variable in our regression (i.e., we regress the female-agency association on the target, the period, the target X period interaction, and the scaled frequency of female referents in the training corpus used to train that word2vec model), our effect of interest did not shrink substantially. The distribution of betas for our effect of interest was largely unchanged (average β = 0.220, t = 470.2, df = 26,999, p < 0.001). The results for the specification curve estimated with this adjustment can be seen in Figure S9.

Figure S9: Specification curve plotting the standardized beta of the interaction term β3 from the regression equation 1.1 when including the frequency of female referents in the regression equation



Note. Red points plot the original specification curve. Blue points plot the curve when controlling for frequency of female referents.

We present word-level analyses to provide further insight into the specific attributes being conferred to women in target companies. To do so, we extracted the average association between each of the 100 agency words and each of the 20 female words, averaging across word vector size and organization group (target versus propensity-matched). We then calculated the difference between the pre and the post period (post – pre), to identify the agency words that had moved the most towards the meaning of female referents. Table S7 gives the complete list of agency words, ordered by the extent to which their cosine similarity with female words increased in the target companies, with an additional column showing how the cosine similarity between that word and female words changed in the propensity matched companies.

Table S7. The change in cosine similarity pre-post for each agency word with the gender dictionary of size 20, for both target and propensity matched firms.

Word	Change in Target Firms	Change in Propensity Matched Firms
direct	0.00265	-0.000291
determined	0.002133	0.001046
independence	0.001699	-0.000194
active	0.001545	-0.002162
autonomy	0.001528	0.000147
tough	0.001406	-0.000704
status	0.001404	-0.000296
original	0.001311	0.001558
resilient	0.001244	-0.00013
confidence	0.001164	-0.001136
ambition	0.001049	-0.000496
competitiveness	0.001025	0.00038
decisive	0.000995	-0.00009
persistent	0.000793	-0.000439
assertive	0.000789	0.00009
competent	0.000749	0.00052
analytical	0.000727	0.000093
effective	0.000721	0.001182
energetic	0.00072	-0.001787
rational	0.00063	-0.000362
bold	0.000622	-0.0004
outspoken	0.000603	-0.000471
competitive	0.000601	-0.001943
ambitious	0.000595	-0.000076
brilliant	0.000589	-0.000278
clever	0.000589	-0.000076
imaginative	0.000545	-0.000175
powerful	0.000535	-0.000271
educated	0.000526	-0.000088
dominant	0.000508	-0.000063
diligent	0.000466	-0.000134
resourceful	0.00045	-0.00009
ingenious	0.000447	-0.000113
superiority	0.0004	-0.000519
vigorous	0.000394	0.000091
autonomous	0.000365	-0.000334
skillful	0.000362	0.000211
practical	0.000348	-0.000437

unwavering	0.000345	0.000036
realist	0.000334	0.000767
perceptive	0.000313	-0.000074
hardworking	0.000302	0.000298
competence	0.0003	-0.000461
freedom	0.000285	0.000529
adamant	0.000265	0.000126
assertiveness	0.000249	-0.00001
individualistic	0.000232	-0.000731
insightful	0.000224	0.000088
smart	0.000209	-0.000338
convincing	0.000208	-0.000259
creative	0.000169	-0.001071
command	0.000151	0.000137
intelligent	0.000146	-0.000838
industrious	0.000106	0.000048
organized	0.000096	-0.000725
meticulous	0.00005	0.000207
proud	0.000034	0.000131
achievement	0.000028	-0.001229
knowledgeable	0.00001	0.000332
assured	0.000002	0.000263
cunning	-0.000011	-0.000004
logical	-0.000069	-0.000001
brave	-0.000228	-0.000091
assert	-0.000259	-0.00065
independent	-0.000264	0.000141
capable	-0.000302	-0.000147
daring	-0.000366	-0.00034
exploration	-0.000382	-0.00011
sharp	-0.000542	0.000015
capability	-0.000546	-0.000588
dynamic	-0.00061	-0.001281
leader	-0.000713	-0.001112
confident	-0.000925	-0.00065
intellectual	-0.001097	-0.000974
serious	-0.001106	-0.000893
recognition	-0.001411	-0.001511
power	-0.001489	-0.000415
fast	-0.001505	-0.001548
efficient	-0.001564	-0.001519
aggressive	-0.001637	-0.000775
strong	-0.002262	-0.001906
unique	-0.003598	-0.002428
individual	-0.004992	-0.003571

Aligning vector spaces with orthogonal Procrustes transformation.

One open question regarding the nature of our effect is whether it is driven by the movement of the gender vectors, the agency vectors, or both. This could entail studying changes in both gender and agency vectors for different kinds of firms (those that did hire female CEOs, and their associated propensity matches). However, the word vectors from the “pre” and “post” models cannot simply be compared, due to them being located in different vector spaces. In order to address this misalignment, we used orthogonal Procrustes to project the vectors estimated for the “post” period into the same space as the “pre” period, which allowed us to estimate the extent to which each vector had moved between the periods pre and post hire. We describe the details of this analysis below.

For the sake of simplicity, we focused this additional analysis on the 300-dimension word vectors. For each of our 33 organizations (11 target, and 22 propensity matches), we transformed the word vector model estimated with the text in the period post-hire, projecting this vector space into the vector space of the model estimated with the text pre-hire. We performed this alignment process only for the relevant agency and female words (as opposed to the entire vector space). To capture movements between pre and post periods, we finally computed the spatial distance between each word’s vector in the pre-hire model and its vector in the post-hire model.

Intuitively, the orthogonal Procrustes method finds the set of transformations (e.g., rotation, scaling, translation) that best superimposes one shape onto another. If the shapes perfectly match, they will perfectly coincide after the Procrustes Superimposition. If they do not perfectly match, one can then compare the overall difference between the two shapes (the Procrustes distance). For our specific application, one can think of each of the focal word vectors as a point denoting the vertex of a shape. The orthogonal Procrustes method maintains the colocation of the post-hire vectors relative to each other, but attempts to orient the word vectors to be aligned with the pre-hire vectors. Any residual deviation of the post-hire vectors from their pre-hire analogues can be interpreted as the extent to which that word vector moved between the periods.

We first study the changes for agency and female vectors just for target firms that hired female CEOs. For each of the 102 agency and female words that occurred in our text corpus, we computed the cosine spatial distance between the pre and post vectors that corresponded to that word. We found that both agency ($t = 26.00$, $df = 912$, $p < 0.001$) and female vectors ($t = 10.78$, $df = 219$, $p < 0.001$) had an average cosine distance that differed significantly from 0. In other words, we found evidence that both agency and female words changed in meaning between the periods pre and post. We find that the agency vectors moved significantly more than the gender vectors ($t = 4.35$, $df = 399.4$, $p < 0.001$). This is as expected, given that the concept of gender should be more constant than the concept of agency and leadership.

Our next analysis compared whether target companies might simply be associated with less stable definitions of leadership and gender – as measured by the disparity in the movement of agency/female vectors between target and propensity matched companies. In contrast to this alternative explanation, we found that the average spatial distance between agency vectors pre-post was actually higher for propensity matches (M = 0.00103) than target organizations (M = 0.000891), and that this difference was statistically significant ($t = 3.03$, $df = 2306.9$, $p = 0.002$). Further, there was no significant difference in the extent to which the female vectors changed ($M_{target} = 0.000605$, $M_{prop\ match} = 0.000616$, $p = 0.867$). In other words, whilst the female-agency association moved by more in the target organizations, this was not a consequence of having a less stable concept of leadership. The agency vectors actually moved more in the propensity matched organizations. This suggests that it was not the sheer magnitude of the change of the meaning of agency and gender in vector space that drove our effects, but rather the specific nature of the changes, whereby organizations viewed women and agency as more similar concepts.

Finally, looking at the 10 agency words that changed most in meaning pre-post in the target organizations (Table S8) and cross-referencing them with results of Table S7 (the words whose association with female words increased the most pre-post for target organizations), we find 4/10 overlap. This supports the idea that the meaning of these words might have changed in part as a result of moving closer to the meaning of being female. We present the complete list of agency and gender words and their cosine distances pre-post for both target and propensity matched organizations.

Overall, the results of this additional analysis help to explain the meaning of the results presented in the main manuscript. First, we find evidence that the meaning of agency changed more in vector space pre-post than the meaning of being female. This supports the suggestion that companies project that the qualities that are conducive to successful leadership (agency) become more similar to the meaning of being female, rather than what it is to be a woman having to change. Second, we do not find evidence for greater instability of the word vectors in the target firms in general. This implies that the observed effects (where we see greater similarity change for the meaning of agency and being female) is driven by a specific, directed change, where the two concepts are seen as more similar, rather than greater change in general.

Table S8. The cosine distance between each word vector estimated in the pre and post period, on average for target and propensity matched organizations.

Word	Change in Target Organizations	Change in Prop Matches
individual	0.003747	0.004395
strong	0.003252	0.003907
effective	0.003090	0.003151
determined	0.002921	0.003330
competitive	0.002906	0.003180
her	0.002798	0.002685
direct	0.002609	0.003034
power	0.002508	0.002975
independent	0.002444	0.002964
original	0.002003	0.002283
active	0.001996	0.002304
unique	0.001921	0.002382
capability	0.001904	0.001812
status	0.001763	0.001661
she	0.001506	0.001508
recognition	0.001415	0.001616

dynamic	0.001392	0.001660
mrs	0.001311	0.000712
aggressive	0.001277	0.001156
achievement	0.001169	0.001580
autonomous	0.001127	0.000578
efficient	0.001124	0.001181
independence	0.001095	0.001202
powerful	0.001016	0.001079
leader	0.000956	0.001033
confidence	0.000944	0.001038
confident	0.000944	0.000976
competitiveness	0.000909	0.000903
intellectual	0.000900	0.000934
fast	0.000798	0.001034
tough	0.000792	0.001032
smart	0.000772	0.001985
creative	0.000722	0.000532
practical	0.000709	0.000752
women	0.000706	0.000771
capable	0.000677	0.000697
daring	0.000672	0.000328
sharp	0.000639	0.000615
intelligent	0.000616	0.001346
proud	0.000613	0.000678
assert	0.000613	0.000600
rational	0.000606	0.000632
serious	0.000595	0.000713
resilient	0.000586	0.000559
imaginative	0.000550	0.000340
organized	0.000522	0.000625
dominant	0.000522	0.000716
unwavering	0.000510	0.000564
freedom	0.000486	0.000494
command	0.000484	0.000517
competent	0.000476	0.000551
persistent	0.000474	0.000501
bold	0.000473	0.000658
decisive	0.000468	0.000467
assured	0.000465	0.000551
superiority	0.000454	0.000473

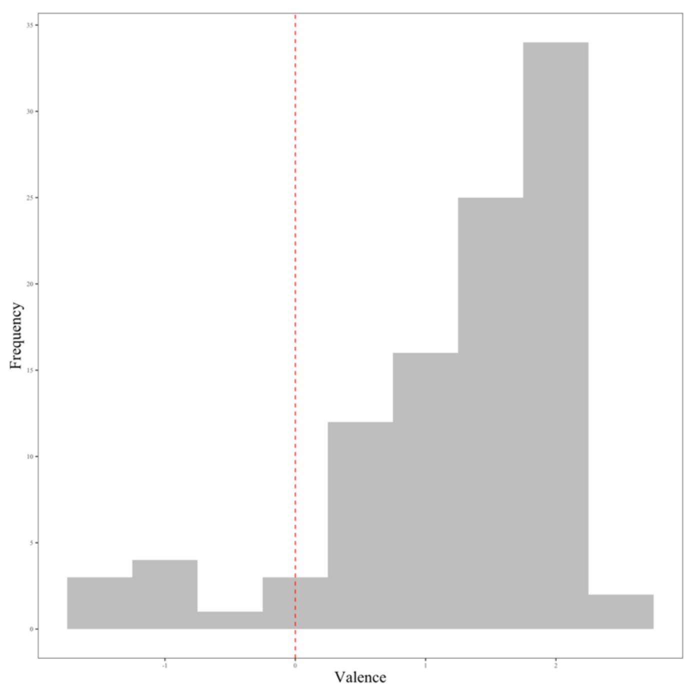
diligent	0.000453	0.000536
logical	0.000447	0.000496
woman	0.000445	0.000488
herself	0.000436	0.000558
female	0.000432	0.000545
females	0.000431	0.000449
ambition	0.000428	0.000526
resourceful	0.000428	0.000782
cunning	0.000423	0.000468
ambitious	0.000419	0.000497
knowledgeable	0.000412	0.000419
daughter	0.000408	0.000353
exploration	0.000402	0.000383
analytical	0.000398	0.000953
brave	0.000398	0.000344
brilliant	0.000393	0.000797
skillful	0.000390	0.000409
educated	0.000377	0.000380
perceptive	0.000376	0.000438
autonomy	0.000375	0.000453
vigorous	0.000373	0.000423
mothers	0.000373	0.000371
competence	0.000362	0.000383
energetic	0.000361	0.000963
ingenious	0.000357	0.000354
feminine	0.000352	0.000384
mother	0.000348	0.000385
meticulous	0.000348	0.000424
daughters	0.000346	0.000444
realist	0.000343	0.000717
outspoken	0.000341	0.000429
industrious	0.000338	0.000388
femininity	0.000331	0.000411
sisters	0.000329	0.000366
hers	0.000327	0.000401
adamant	0.000327	0.000430
maternal	0.000322	0.000328
sister	0.000321	0.000367
hardworking	0.000312	0.000405
convincing	0.000306	0.000463

girl	0.000303	0.000438
assertiveness	0.000301	0.000404
individualistic	0.000298	0.000619
assertive	0.000297	0.000390
clever	0.000292	0.000409
insightful	0.000280	0.000378
girls	0.000280	0.000364

Dividing agency dictionary by valence and type of agency

To test whether the increase in the female-agency association in target firms was driven by the positive or negative side of agency, we split our dictionary of 100 agency related words by the valence ratings we collected when validating our dictionaries. 92 words had a valence rating above 0 (on a scale of -3 to 3), and 8 words had a valence rating below 0 (see Table S4a). We estimated two additional specification curves following the exact same procedure as before, but using either positive or negative valence agency words as the theory dictionary input (**Fig. 4a-b**). See Figure S10 for the distribution of valence ratings of our agency dictionary.

Figure S10: A histogram of the valence ratings of our agency dictionary.



Secondly, we aimed to better understand which nuances of agency were responsible for the increase in the association. Competence and dominance are two widely theorized components of agency (27, 28). Based on this distinction we constructed dictionaries of 35 competence-related versus 29 dominance-related words (e.g., ‘capability’ versus ‘power’), by collecting ratings of the association between each word in our 100-word agency dictionary and competence / dominance using Amazon's Mechanical Turk platform. We recruited 55 participants, with an average age of 42.7 years and 51% women. Using these ratings, we defined a 'dominant agency' dictionary by requiring an average dominance rating of over 5 and an average competency rating of below 4.5, and vice versa for our 'competency dictionary'. See Figure S11 for the dominance and competency ratings plotted as a scatter graph, and Table S9 for the full list of dominant and competent agency words. The negative correlation between these ratings ($corr(X,Y) = -.74$, $p < 0.001$) lends some support for the discriminant validity of these two dimensions of agency. We estimated two final specification curves using these 'competency' and 'dominance' agency dictionaries, in order to verify if there were particular aspects of agency that became more closely associated with women after the event of hiring a female CEO.

Figure S11: A scatter graph plotting our competence and dominance ratings against each other.

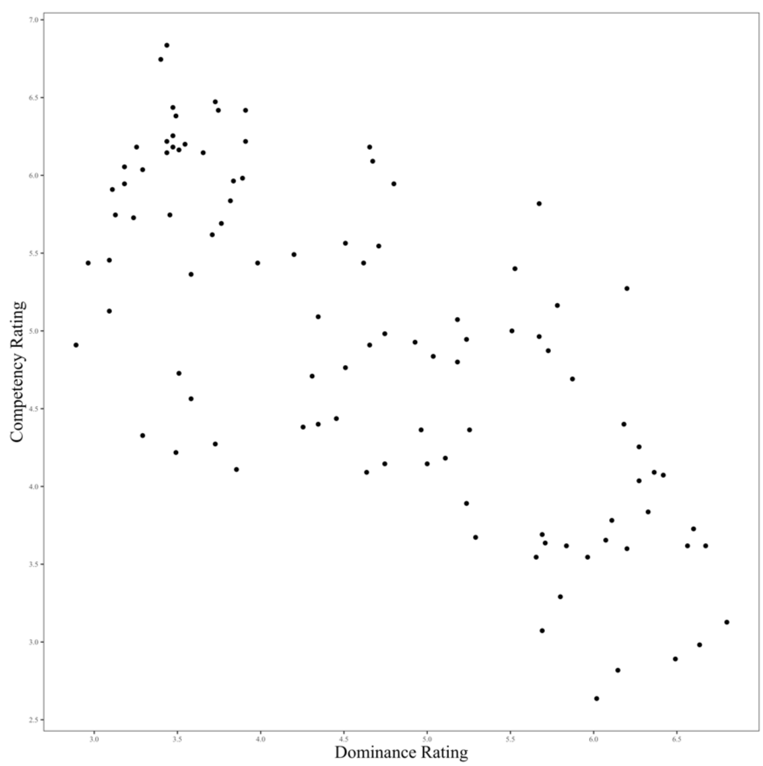


Table S9: Competence and dominance agency dictionaries.

Dictionary	Word	Dominance Rating	Competence Rating
dominant	dominant	6.80	3.13
dominant	command	6.67	3.62
dominant	aggressive	6.64	2.98
dominant	powerful	6.60	3.73
dominant	power	6.56	3.62
dominant	bossy	6.49	2.89
dominant	superiority	6.42	4.07

dominant	competitive	6.36	4.09
dominant	assertive	6.33	3.84
dominant	strong-willed	6.27	4.04
dominant	competitiveness	6.27	4.25
dominant	assertiveness	6.20	3.60
dominant	strong	6.18	4.40
dominant	egoistic	6.15	2.82
dominant	assert	6.11	3.78
dominant	bold	6.07	3.65
dominant	egocentrist	6.02	2.64
dominant	outspoken	5.96	3.55
dominant	tough	5.84	3.62
dominant	adamant	5.80	3.29
dominant	status	5.71	3.64
dominant	direct	5.69	3.69
dominant	self-important	5.69	3.07
dominant	daring	5.65	3.55
dominant	unwavering	5.29	3.67
dominant	cunning	5.25	4.36
dominant	proud	5.24	3.89
dominant	vigorous	5.11	4.18
dominant	recognition	5.00	4.15
competent	competence	3.44	6.84
competent	competent	3.40	6.75
competent	capability	3.73	6.47
competent	knowledgeable	3.47	6.44
competent	capable	3.91	6.42
competent	skillful	3.75	6.42
competent	skillfulness	3.49	6.38
competent	smart	3.47	6.25
competent	effective	3.91	6.22
competent	brilliant	3.44	6.22
competent	intellectual	3.55	6.20
competent	intelligent	3.47	6.18
competent	logical	3.25	6.18
competent	efficient	3.51	6.16
competent	resourceful	3.65	6.15
competent	educated	3.44	6.15
competent	well-organized	3.18	6.05
competent	organized	3.29	6.04
competent	hardworking	3.89	5.98
competent	well-disciplined	3.84	5.96
competent	rational	3.18	5.95
competent	insightful	3.11	5.91
competent	industrious	3.82	5.84
competent	practical	3.13	5.75
competent	ingenious	3.45	5.75
competent	analytical	3.24	5.73
competent	diligent	3.76	5.69

competent	perceptive	3.71	5.62
competent	sharp	4.20	5.49
competent	meticulous	3.09	5.45
competent	clever	3.98	5.44
competent	creative	2.96	5.44
competent	self-control	3.58	5.36
competent	realist	3.09	5.13
competent	resilient	4.35	5.09

Raw cosine similarities. As part of our dictionary split analysis that aims to provide explanation of the mechanisms driving our effect, we report the raw cosine similarities between our gender dictionaries and our agency dictionaries. This analysis requires clarification of several trivial details that are not included in the manuscript due to space constraints. To calculate these average cosine similarities at a *dictionary-level*, we averaged across companies (only the 33 target companies, excluding the 6 additional companies) and word vector sizes (100, 300, 500). We used only the gender dictionary of size 20 for parallelism with the word level analyses (where we analyze the average cosine similarity between each agency word and the gender dictionary of size 20).

Note that some of the words we specified using a theoretical basis either 1) did not occur more than once in the training corpus for the transfer learning pre-trained model, or 2) did not occur in specific firms' updating corpora. A word had to occur more than once for our algorithm to create a word vector in the base model, and had to occur a positive number of times to be updated in a specific model. Neither of these bias the results of our analysis – if a word did not occur in the pre-trained model, no word vector was created, and hence this did not contribute to the findings of our analysis. If a word did not occur in a specific firm's updating corpus, the corresponding word vector remained unchanged.

We account for missing word vectors in our calculation of the average cosine similarity between our gender dictionaries and agency dictionaries by subtracting one from the denominator of the calculation (sum / number of pairs) for each time the code is unable to compute this cosine similarity. This means that we get accurate estimates of the average cosine similarity between the gender and agency dictionaries. 83/100 of the agency words occurred in the pre-training corpus, and the average cosine similarity between the agency dictionary of size 100 and the gender dictionary of size 20 can be recovered by averaging the similarities between each word and this gender dictionary. As for those firms for which specific words did not update, the code functions in a manner analogous to grand mean imputation – for that firm, the word vector for that word remains the same as the word vector from the pre-trained model, which serves as a prior.

Out of our positive agency dictionary of 92 words, 79 were matched in the pre-training data. Out of our negative agency dictionary of 8 words, 4 were matched in the pre-training data. Hence, a simple weighted average of the reported raw cosine angles will be inaccurate, due to some of these words not occurring in the pre-training data. When using the adjusted lengths of dictionaries, the aggregate figure for each cosine similarity can be recovered from both its constituent dictionaries (e.g., positive and negative) and the word-level value (e.g., the 83 agency words for which vectors exist).

Dictionary Choices and Robustness

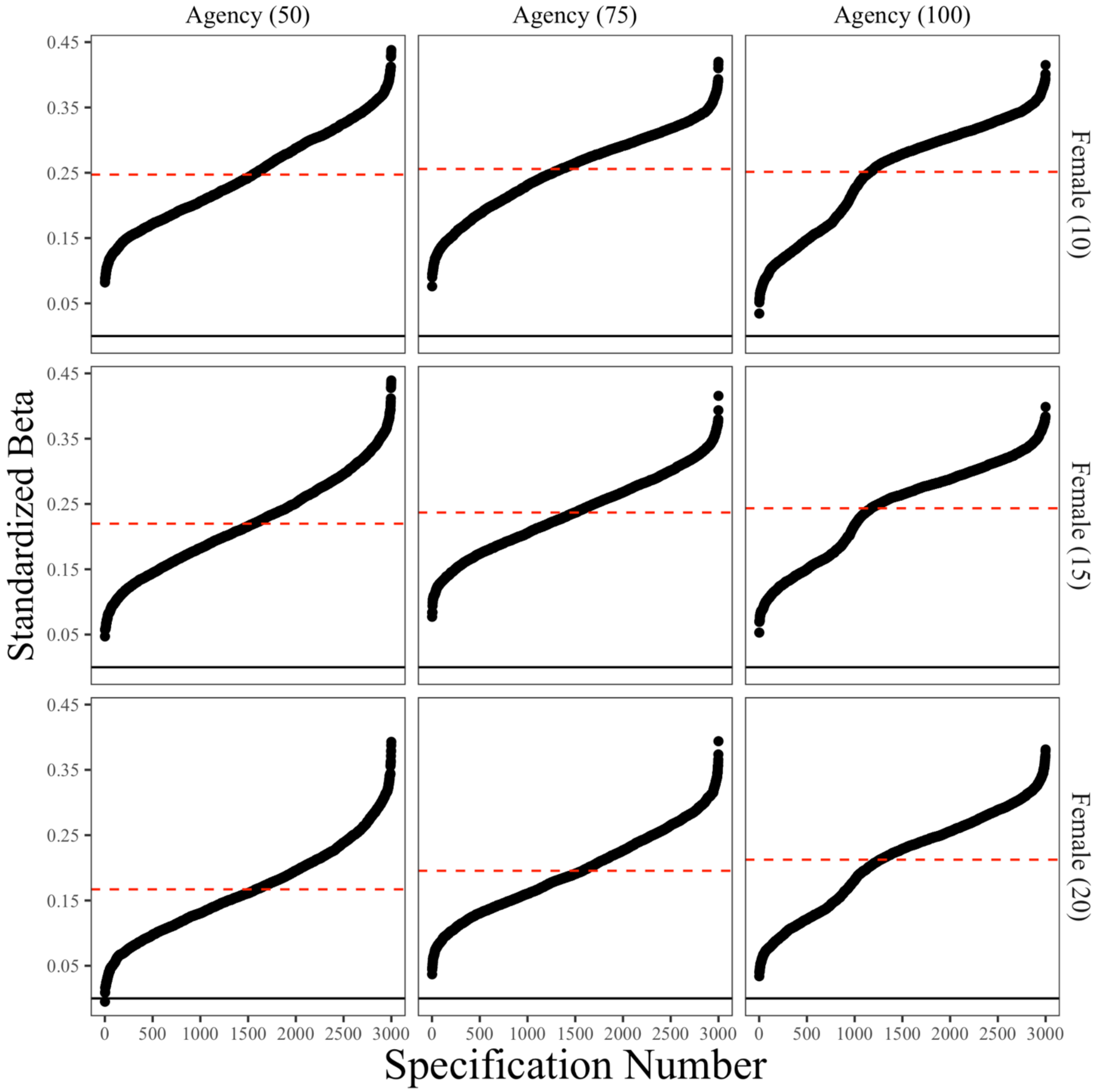
Determining which words best captured the concepts of gender and agency involved several considerations. For agency, we discuss our selection process in full earlier in the supplement. For the gender dictionaries, we wanted to capture the meaning of being a woman in a broad sense, but believed that female pronoun references (e.g., “she”) were the most likely candidates to provide insight regarding the gendering of language. Such words are used to refer to all women, and thus do not face the potential to be biased by perceptions of a particular woman. For example, a specific female title such as “duchess”, whilst gendered, is likely closely coupled to a specific individual, thus limiting the generality of any conclusions that can be reached. Further, for a gender word to be meaningfully updated in a particular organization’s corpus, it had to occur relatively frequently.

In order to test the robustness of our conclusions to different specifications of gender dictionaries, we conducted a series of additional analyses. First, we show how our conclusions vary when using different sized gender dictionaries. Second, we contemplate the addition of other female referents to our gender dictionaries, and flesh out the rationale behind the main gender dictionaries that we used. Finally, we replicate our main effect with the inclusion of the word “ms” in our female dictionary.

Faceted by dictionary size. Whilst the effects of gender and agency dictionary size are constant within specification, and hence do not bias the results of our regressions that compare the relative change in the female-agency association between firms that did and did not hire female CEOs, it is of interest to understand how the choices regarding these dictionaries affect the overall effect size. To this end, we present the data from the specification curve **Fig. 3** in the manuscript, faceted by the three gender dictionary sizes and three agency dictionary sizes (Figure S12). In the 3x3 matrix of possible gender and agency dictionary configurations, the mean value of the coefficient β_3 varied from 0.167 to 0.256. We see that the effect size decreases as the gender dictionary gets larger (a potential consequence of more noise associated with less targeted gender dictionaries) and is generally slightly larger for the agency dictionary of size 75, relative to 100, and 100 is slightly larger relative to size 50.

This analysis provides some suggestion that the effect is stronger for a more focused set of gender referents. The female dictionary size 10 contained the words ['mrs','hers','her','she','women','woman','female','herself','females','feminine'], whereas the size 15 added the words ['femininity','daughter','daughters','mothers','mother'], and size 20 further added ['maternal','sister','sisters','girl','girls']. This pattern of results is face valid – whereas the smaller dictionary’s words are tightly coupled to references to women in the organization, the wider sets consider aspects of womanhood that also capture family dynamics. This may explain the slight reduction to the effect size when using the larger gender dictionaries.

Figure S12: Specification curve plotting the standardized beta of the interaction term β3 from the regression equation 1.1 faceted by size of agency dictionary and size of gender dictionary.



Additional female referents. We additionally considered a wider set of gendered references after our pre-registration period. This involved considering explicitly gendered terms for which male and female equivalents exist. For example, "businesswoman" and "businessman". However, when calculating the relative frequency of such words, we found them to be very rare. Table S10 includes the frequency of occurrence of a range of gendered references. Due to this rarity (or complete non-occurrence) of these words in our updating corpora, we did not further pursue using such words to capture gendered references. It is possible (and likely) that in other contexts, such as references to female leaders in the media, that such referents would be helpful in broadly capturing the meaning of being female.

Table S10: Count of expanded female referents in the updating corpora.

Word	Count
chairwoman	1
waitress	1
anchorwoman	0
businesswoman	0
committeewoman	0
congresswoman	0
councilwoman	0
gentlewoman	0
heroine	0
laywoman	0
madwoman	0
policewoman	0
saleswoman	0
servicewoman	0
spokeswoman	0
stateswoman	0
womankind	0

The female referent "ms". In constructing gender dictionaries, we aimed to identify words that were both frequently used and unambiguously referred to females. In pursuing these joint goals, the reference "ms", as in "Ms. Smith" poses a challenge to the second condition. Whilst this title is often used to refer to women in formal contexts, the two letters "ms" also have many other interpretations. In the following section we offer two analyses aimed at addressing this point.

First, we present a sample of the context words surrounding uses of the word "ms" in our text corpus (Table S11). Whilst there are numerous examples of appropriate uses of the word, there are enough misleading uses to cast doubt on its validity. The fact that "ms" could refer to Mississippi, Mass Spectroscopy, or Multiple Sclerosis is not a problem faced by our other dictionary words. It is difficult to develop a rigorous method of testing the ratio of relevant vs. irrelevant uses of the word because the surrounding context words vary so much. For example, fully specifying the possible definitions of irrelevant uses (e.g., Master of Science, etc.) is insufficient to determine when the two characters ("ms") are being used to denote this meaning. Moreover, Microsoft and Morgan Stanley are organizations that feature in our pre-training corpus, both of which are often referred to as MS. Given that our pre-trained models serve as relatively strong priors, if they have encoded alternative uses of the word "ms" in its word vector, updates from this base may be harder to interpret.

As a result of these considerations, we elected not to augment our pre-registered dictionaries with the referent "ms". However, we did aim to test the robustness of our conclusions to this design choice, and thus also replicated our main effect when including "ms" in our female gender dictionary (Figure S13). We outline these analyses in full below.

Table S11. Context words surrounding uses of the word "ms".

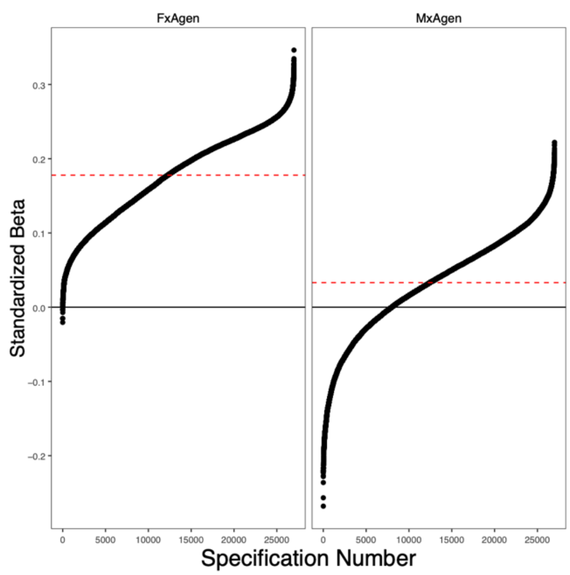
Example use
1. community care of mississippi inc ms amerigroup community care of new mexico inc nm

2. complementing the companys technology leadership in mass spectroscopy ms some of our products support a complete protein analysis workflow in ms analysis
3. products are indicated for the treatment of patients with relapsing forms of multiple sclerosis ms

Below, we include our main result re-estimated including the word "ms" in the female gender dictionary. We count "ms" as an equivalent word to "mrs", so both are included in the dictionary size 10, 15, and 20 (which are then technically 11, 16, and 21). The denominator of our averaging procedure is adjusted accordingly.

Main specification curve. Providing support for our hypothesis, the findings show that the pre-post change in female-agency association is significantly larger for organizations that hired a female CEO (average β3 = 0.178, SD(β3) = 0.061). With the vast majority of standardized regression coefficients being positive (99.967%), the distribution of coefficients is significantly larger than zero (CI_{95} = [0.1771, 0.1786], t = 476, df = 26,999, p < 0.001). An independent t-test comparing the distributions of beta coefficients for the female and male regression models confirms that the standardized interaction effects on the female-agency association were significantly larger than those for the male-agency association (*t* = 1,913, *df* = 26,999, *p* < 0.001).

Figure S13: Main specification curve with modified gender dictionary.



Board Level Analysis

The dataset. In our text data sample (including 10-K and DEF 14-A filings, and transcripts of conference calls) we had data relating to 690 unique organizations in the period 1st January 2009 to 31st December 2018. There were an additional 6 organizations that we removed due to inconsistencies between their identification via CIK and ticker. We randomly selected half of these eligible organizations (345) and used their full text data to pre-train a word2vec model with 100 dimensional vectors. We then estimated 8 word2vec models for each of the other 345 organizations, using 3-years of data for each model. For example, we estimated a model for Apple using just documents from 1st January 2009 to 31st December 2011, and another model using documents from 1st January 2010 to 31st December 2012, and so on up until the period 1st January 2016 to 31st December 2018. This sliding window allowed us to obtain longitudinal estimates of linguistic stereotypes at a company level, whilst affording more reliable estimates than single year periods would. We chose to use only 100 dimensional word vectors to reflect the reduced complexity of the data as a result of smaller amounts of text per model.

We recovered director information from the Institutional Shareholder Services (ISS) database in Wharton Research Database Services (WRDS) and aggregated it to recover organizational-level data regarding the gender diversity of corporate boards of directors. This data captures the number of women on each organization's board of directors, and the total number of people on the board of directors. This allowed us to express each board's female representation as a single value – the proportion of female directors. We averaged this yearly data across the periods that corresponded to the word2vec models and merged the two datasets.

After removing periods for which we had no text data, and thus could not update the word2vec model, and periods where director information was not available, we were left with a dataset of 2,286 observations from 328 different organizations (an average of 6.97 observations per organizations). For each observation, we measured the association between women and agency, women and communality, men and agency, and men and communality, as well as the proportion of women on the board of directors. We simplified our analysis to focus on the associations between our restricted gender dictionaries (10 words each) and the full agency and communality dictionaries (100 words each). We standardized each linguistic variable to put them on similar scales.

Modelling procedure. We aimed to measure the effect of changes in women's representation on boards of directors on gender stereotypes expressed in language. However, we were concerned with the possibility of bidirectional causality. Appointing more women to the board of directors could lead to mitigation of gender stereotypes, but changes to linguistic stereotypes could also precipitate changes in female representation. Considering this concern, we opted to analyze our data using panel vector autoregression (29). Panel vector autoregression is a modelling method, used when more than one variable is endogenous that does not require specifying relationships between endogenous variables (29, 30). Consider a simple model measuring the relationship between the female-agency association (F_t) and proportion of women on the board of directors (P_t) over time (t). In a two-period simplification, the model would analyze the stability relationships within each variable (i.e., F_1 and F_2 and P_1 and P_2) as well as the association between a lag of the female-agency association and the proportion of women (F_1 and P_2) and a lag of the proportion of women and the female-agency association (P_1 and F_2). This modelling approach allows one to extricate the effect of F on P from the effect of P on F.

Parameter specification. Using panel vector autoregression requires the specification of several parameters in order to estimate the model. We delineate the empirical support and theoretical rationale behind each of our choices in the following section. First, we opted to use a generalized method of moments (GMM) estimator (31) rather than OLS-based regression, due to the Nickell bias (32) not disappearing asymptotically in the number of units of observation (organizations) for a fixed number of time periods (29).

Within the GMM estimator, we needed to determine several parameters. These parameters included; which of our variables were endogenous, the number of lags of our endogenous variables, the transformation of lags (first difference or forward orthogonal transformation), whether to collapse the first difference GMM moment conditions, and whether to use system instruments or not. We first discuss aspects of the estimation (the transformation of lags, collapsing moment conditions, and system instruments), before discussing different design choices concerning endogeneity, and their implications for our results.

The transformation of lags (first difference versus forward orthogonal transformation) will yield the same estimation results in a balanced panel when the full set of available instruments is used (29). The forward orthogonal transformation was proposed by Arellano and Bover (33) to minimize data losses due to data gaps. Given that our panel is imbalanced (there are different numbers of observations for different organizations, and some missing periods as a result of there being no new text data), we opted to use the forward orthogonal transformation to avoid data losses. Phillips (34) shows using Monte Carlo simulations that estimators based on forward orthogonal deviations have better finite sample properties than estimators based on differencing.

Collapsing the GMM moment conditions yields several benefits. First, this reduces concerns regarding instrument proliferation (35). Based on asymptotic theory, it also increases the likelihood of an estimator being consistent with a smaller sample (36). Given our relatively modest panel size, this is a desirable property. Finally, reducing the number of moment conditions also increases computational efficiency (29). Considering these factors, we elected to collapse these conditions. For further details of this process, see Sigmund and Ferstl (29).

The system based GMM estimator performs better than the FD GMM estimator, as the additional instruments remain good predictors of the endogenous variables even when the series are very persistent (37). However, the system GMM estimator requires the stronger assumption that the changes in the variables are uncorrelated with the unobserved fixed effects. This assumption is met in a stationary panel vector autoregression model (i.e., when all eigenvalues of the PVAR polynomial are less than 1). We tested whether our models were stationary, and if they were, opted to use these system instruments in our estimation process to achieve superior finite sample properties. We delineate these analyses in full in the following section.

Finally, we had to choose which of our variables to include in our model, which variables would be endogenous, and the number of lags to include of each endogenous variable. We compared three different sets of models, with the following sets of variables:

1) Model 1
  - **a.** Endogenous: proportion of female directors, female-agency association, female-communality association
  - **b.** Exogenous: NA

2) Model 2
  - **a.** Endogenous: proportion of female directors, female-agency association, female-communality association
  - **b.** Exogenous: male-agency association, male-communality association

3) Model 3
  - **a.** Endogenous: proportion of female directors, female-agency association, female-communality association, male-agency association, male-communality association
  - **b.** Exogenous: NA

We first tested whether each of these models was stationary with 1 lag, in order to ascertain whether to use a system based GMM estimator. We next followed the procedure of Andrews and Lu (38) to choose the optimal number of lags for each variable specification. Finally, we present and interpret estimates of each set of models.

Stationarity tests

1) Model 1: satisfied Eigenvalue stability condition (i.e., was stationary) when not using system based GMM estimator (max = 0.781), but not when using system estimator (max = 1.341).
2) Model 2: satisfied Eigenvalue stability condition (i.e., was stationary) when not using system based GMM estimator (max = 0.763), but not when using system estimator (max = 1.339).
3) Model 3: satisfied Eigenvalue stability condition (i.e., was stationary) when not using system based GMM estimator (max = 0.756), but not when using system estimator (max = 1.349).

Given the results of these tests, we did not use the system GMM estimator for any of our model estimations. In summary, we used a generalized method of moment estimator with forward orthogonal transformations, collapsed GMM moment conditions, and no system based instruments.

Model and moment selection criteria (MMSC). Following Andrews and Lu's (38) recommendation, we selected the number of lags that minimized the BIC (Bayesian information criterion). In each case, this was 1 lag. This also achieved the goal of facilitating greater interpretability and minimizing data loss. With longer numbers of lags shorter panels (e.g., firms for whom we only had 2 or 3 observations) must be excluded.

Table S12. Model and moment selection criteria

Model 1:

	MMSC		
Lags	BIC	AIC	HQIC
1	-300.5	-25.0	-137.4
2	-288.6	-38.8	-141.9
3	-279.9	-56.5	-149.9

Model 2:

	MMSC		
Lags	BIC	AIC	HQIC
1	-332.2	-35.0	-156.2
2	-319.4	-48.8	-160.4
3	-305.7	-62.5	-164.2

Model 3:

	MMSC		
Lags	BIC	AIC	HQIC
1	-940.1	-156.7	-476.2
2	-909.2	-180.6	-481.1
3	-857.3	-187.1	-467.4

Following this rigorous procedure to choose empirically validated parameters, we estimated three PVAR models, using each of the three possible variable specifications that we proposed (Table S13). We present the full model estimates for Models 1-3 below. In each case, the brackets indicate the p-value associated with the coefficient estimate.

Table S13. Panel vector autoregression models

Number of groups 328

Proportion Women (PW)
Female-agency (FA)
Female-communality (FC)
Male-agency (MA)
Male-communality (MA)

Model 1	PW (t)	FA (t)	FC (t)
PW (t-1)	0.765 (<0.001)	2.794 (0.039)	1.517 (0.329)
FA (t-1)	0.006 (0.007)	0.295 (0.001)	0.065 (0.456)
FC (t-1)	-0.005 (0.022)	-0.065 (0.373)	0.236 (0.009)

Model 2	PW (t)	FA (t)	FC (t)
PW (t-1)	0.762 (<0.001)	2.270 (0.084)	2.307 (0.105)
FA (t-1)	0.007 (0.006)	0.239 (0.003)	0.087 (0.272)
FC (t-1)	-0.006 (0.009)	-0.056 (0.422)	0.147 (0.071)
MA (t)	-0.001 (0.480)	0.282 (<0.001)	-0.247 (<0.001)
MC (t)	0.002 (0.247)	-0.072 (0.127)	0.477 (<0.001)

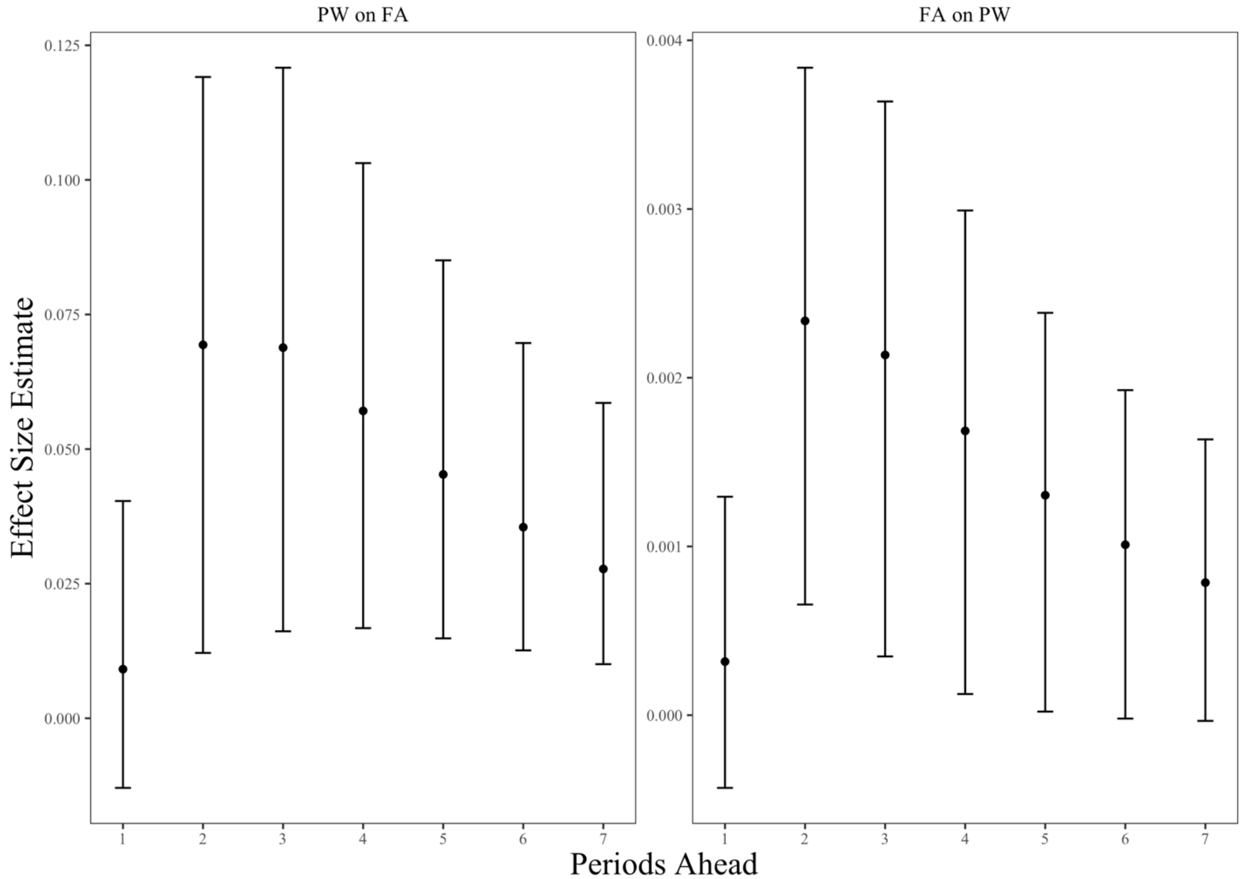
Model 3	PW (t)	FA (t)	FC (t)	MA (t)	MC (t)
PW (t-1)	0.731 (<0.001)	2.771 (0.043)	1.524 (0.308)	1.39 (0.311)	0.825 (0.519)
FA (t-1)	0.006 (0.009)	0.274 (<0.001)	0.064 (0.395)	0.068 (0.392)	0.071 (0.316)
FC (t-1)	-0.004 (0.117)	-0.02 (0.766)	0.234 (0.004)	-0.023 (0.783)	-0.025 (0.744)
MA (t-1)	0.004 (0.131)	0.005 (0.941)	-0.041 (0.573)	0.287 (0.001)	-0.125 (0.116)
MC (t-1)	-0.005 (0.072)	-0.075 (0.290)	0.017 (0.804)	0.104 (0.263)	0.481 (<0.001)

The effect of the proportion of women on the female-agency association was statistically significant in our primary specification (Model 1; $b = 2.794$, $p = 0.039$). When including the male-agency and male-communality associations as exogenous variables, this fell slightly below the threshold for statistical significance (Model 2; b = 2.270, p = 0.084), but when including all variables as endogenous, this achieved statistical significance again (Model 3; b = 2.771, p = 0.043). Overall, these results are show that there is a positive effect of the proportion of women on an organization's board of directors on the female-agency association in the organization.

Of the sample of organizations that we used to estimate these panel vector autoregression models, 97.3% were unique from our sample of organizations that hired female CEOs in the sample period. When excluding the 2.7% of the firms that overlapped, the effect of PW (t-1) on FA (t) decreased in significance to slightly below the 5% threshold (*b* = 2.597, *p* = 0.068).

Interpreting the effect sizes. In a panel vector autoregression (PVAR) model, because variables depend on each other, individual coefficients provide limited information regarding the system's response to a shock. As a result, impulse response analysis is used to estimate the effect of an impulse to one endogenous variable on another endogenous variable (29). In the follow section, we detail generalized impulse response analysis for Model 1 (Table S13). Using the generalized impulse response function (GIRF) has the advantage over other analytical approaches (i.e., orthogonal impulse response functions (OIRF)) in that it is unaffected by the ordering of variables (39). The GIRF measures the response of endogenous variables to a one standard deviation shock to another endogenous variable. For Model 1, we present a 7-period GIRF (Figure S14), with 95% confidence intervals computed using the bootstrapping procedure defined in Kapetanios (40). This shows the dynamic response of the focal endogenous variables to a one standard deviation shock to a variable that is implied by our model coefficients. We also provide interpretation of the total cumulative effect sizes implied by this analysis, in light of the coding of our variables.

Figure S14. Generalized impulse response function for Model 1



Note. Error bars indicate 95% confidence intervals

The left panel of Figure S14 shows the development of the effect of a change to the proportion of women on a board on the female-agency association, whereas the right panel shows the effect of a change to the female-agency association on the proportion of women on the board. The effects follow similar trajectories, with the largest one-period increase coming in the second period after the shock, before decaying in the following periods. However, the effect of the proportion of women changing on the female-agency association is much larger than the effect of the female-agency association changing on the proportion of women. Across the entire period, a one-SD shock to PW[1] is associated with FA increasing by 0.313 SD. A one SD shock to FA is associated with PW increasing by 1 percentage point. For the effect of PW on FA, 65.3% accumulates within 4 periods of the shock, increasing to 79.8% after 5 periods, and 91.1% after 6 periods.

[1] In the sample, the standard deviation of the proportion of women on a board of directors was 9.15%. The mean number of directors was 10.3. Hence, a 1 SD shock to the proportion of women was roughly equivalent to appointing one woman on the board of directors.

Additional analyses

Longitudinal analysis of CEO data

In addition to our main modelling procedure, we also produced 24 models that collapsed across target and propensity matched firms, but provided greater granularity with regards to the time period. Each of our 11 target firms had a specific hiring date. This date was defined as the central point of the panel. For the 22 propensity matches, they took their hiring date from their associated target firm. We then divided the text for each target firm and its associated propensity matches into 6-month periods oriented with respect to each hiring date. This gave us a total of 12 periods – 3 years pre and post the hiring date in 6 month increments. We then collapsed across the firms to get a single dataset for the target firms and a single dataset for the propensity matched firms in each of these 6 month periods and trained 100-d word2vec models on this data. This was necessary to have sufficient data to obtain reliable model estimates. In sum, the time periods are defined relative to the hiring dates not a single chronological timeline.

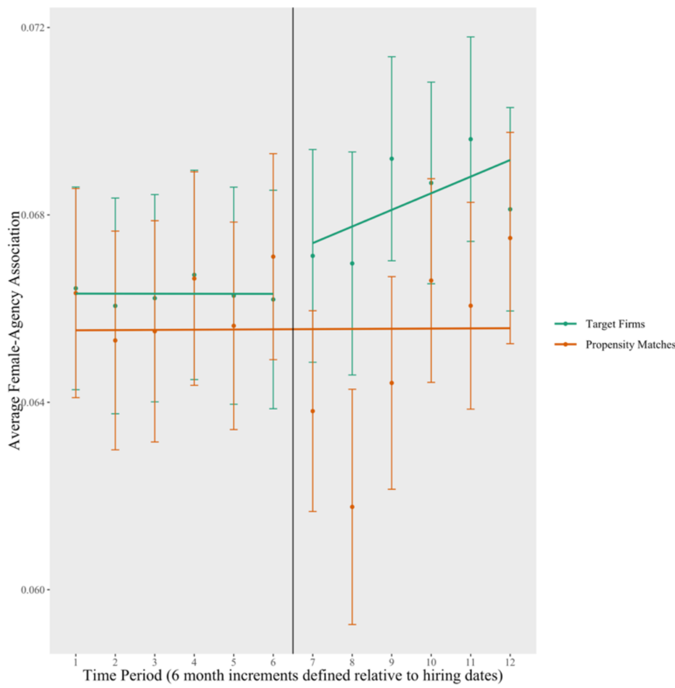
This analysis has several weaknesses. If a firm produced more text data in a particular period, its language change would be overweighted in the model estimates. This, in turn, means that the objective timeline becomes imbalanced, which could potentially confound any trends with wider trends in the ecosystem. For example, if a firm that had an early hire date produced a lot of data in a particular period, whatever natural trends were occurring in that time period (e.g., 2015) might dominate the nuances of trends in post-hire changes in language. Our belief is that the pre/post design we pre-registered and used throughout our main analyses is the best way to trade-off the data requirements of the word2vec algorithm with sufficient granularity and generality to understand the nature of effects. However, acknowledging the desire to better understand the longitudinal nature of the effect, this supplementary analysis provides suggestive evidence regarding the timeline of these effect's progression. In Study 2 in the main manuscript, we provide evidence of how the effects play out over time with a different form of female representation – on boards of directors. We consider this to be more informative, but wish to provide the reader with a comprehensive account of our data.

In Figure S15 we plot the estimates of the female-agency association for each of these groups of firms (target / propensity match) in each of the 12 periods. We include point estimates and error bars indicating + or – 1 standard error (these standard errors are recovered from the 9 separate estimates we have for each measure, when looking across the 3x3 configurations of gender dictionary size and theory dictionary size). We fitted two linear trends to the 6 periods pre and post for the target firms, and a single linear trend to the propensity matches that did not undergo the event of hiring a female CEO.

In the periods pre-hire, the target and propensity matched firms look very similar in terms of their female-agency association. The fitted linear trend for the target firms is at a slightly higher level, but the two are basically flat. We see between periods 6 and 7 (i.e., comparing the 6 months immediately pre and post hire) there is something of a jump in the target firms (and a drop in the propensity matches). The target firms' female-agency association continues to increase, reaching a maximum in period 11 (24-30 months post-hire), before falling slightly back in the final period. This is consistent with the proposal that firms who hire a female CEO are motivated to signal her agency, and that this is not solely an artefact of the hiring event, or short term signaling of their CEO (who happens to be a woman)'s competence. We observe a sharp dip and recovery in the female-agency association for the propensity matches in the period post-hire. This could be a reactance effect to the hiring of female CEOs in the target companies if these propensity matches are direct competitors, but we can only speculate as to the source of this pattern.

Whilst we caveat this analysis with the aforementioned weaknesses – the potential for overweighting specific firms and time periods – it is suggestive that there is an initial effect of hiring a female CEO, and a more persistent upwards trend in the female-agency association that follows.

Figure S15. The progression in female-agency association in 6 month periods for the target and propensity-matched firms.



Case analysis of female-male hires.

We collected data and estimated word2vec models for two companies (Anthem, Inc. and Archer-Daniels-Midland) that hired male CEOs with female predecessors in the sample period. Unfortunately, this number is clearly too small for any meaningful statistical inference. We provide case analysis of these two instances below. Notably, the extremely small sample size makes the analyses highly speculative. First, in Figure S16 we present the average cosine similarities between the female and agency vectors (averaged across gender dictionaries, theory dictionaries, and vector sizes) for each of the two female-male hire firms, and their associated propensity matches. For both companies that replaced a female CEO with a male CEO, we see a decrease in association between women and agency vectors.

However, even with this result, there are two clear reasons for caution; 1) any conclusions from a sample size of two is clearly astatistical, and the larger sample of male-female hires should be considered much more reliable, and 2) without reference to the propensity matches, we do not control for the effect of time / industry trends, that may affect this result. Indeed, the female-agency association decreased in 2/4 of the propensity matches, in line with what appears to be a general trend downwards in the female-agency association that we also observed in the 22 propensity matches from our core sample.

Figure S16. Average cosine similarity between female and agency vectors in periods pre/post for firms that hired male CEOs after female CEOs.

Overall, across our many robustness checks and case analysis of this ancillary sample, we find strong support for the robustness of an effect of hiring a female CEO on the association between female vectors and agency vectors that is not explained merely by talking about the CEO (who happens to be female) more, or a statistical or methodological artefact. The case analysis of this smaller sample perhaps provides a weak signal that replacing a female CEO with a male can lead to backsliding in the progress in changing stereotypes. Future research might examine this change in a larger sample – how do companies gender their language after *replacing* a female CEO? Is progress stable and lasting, or is it the tangible presence of a female leader that drives the need to signal that she is agentic (as are women more broadly)?

References

1. M. Caliendo, S. Kopeinig, Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys* **22**, 31–72 (2008).

2. R. H. Dehejia, S. Wahba, Propensity Score-Matching Methods for Nonexperimental Causal Studies. *Review of Economics and Statistics* **84**, 151–161 (2002).

3. A. E. Abele, B. Wojciszke, Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology* **93**, 751–763 (2007).

4. D. M. Young, C. A. Moss-Racusin, D. T. Sanchez, Implicit agency, communality, and perceptual congruence in couples: Implications for relationship health. *Journal of Experimental Social Psychology* **55**, 133–138 (2014).

5. C. L. Holt, J. B. Ellis, Assessing the Current Validity of the Bem Sex-Role Inventory. 13 (1998).

6. P. D. Trapnell, D. L. Paulhus, Agentic and Communal Values: Their Scope and Measurement. *Journal of Personality Assessment* **94**, 39–52 (2012).

7. L. R. Ramsey, Agentic traits are associated with success in science more than communal traits. *Personality and Individual Differences* **106**, 6–9 (2017).

8. P. Brosi, M. Spörrle, I. M. Welpe, M. E. Heilman, Expressing pride: Effects on perceived agency, communality, and stereotype-based gender disparities. *Journal of Applied Psychology* **101**, 1319–1328 (2016).

9. J. M. Madera, M. R. Hebl, R. C. Martin, Gender and letters of recommendation for academia: agentic and communal differences. *The Journal of applied psychology* (2009) https:/doi.org/10.1037/a0016539.

10. R. L. Schaumberg, F. J. Flynn, Self-reliance: A Gender Perspective on its Relationship to Communality and Leadership Evaluations. *AMJ* **60**, 1859–1881 (2017).

11. N. M. Hale, J. Hewitt, Negative Aging Stereotypes and the Agentic-Communal Dichotomy. *Percept Mot Skills* **87**, 915–918 (1998).

12. V. S. Helgeson, Relation of agency and communion to well-being: Evidence and potential explanations. *Psychological Bulletin* **116**, 412–428 (1994).

13. C. Saragovi, J. Aubé, R. Koestner, D. Zuroff, Traits, Motives, and Depressive Styles as Reflections of Agency and Communion. *Pers Soc Psychol Bull* **28**, 563–577 (2002).

14. A. E. Martin, K. W. Phillips, What “blindness” to gender differences helps women see and do: Implications for confidence, agency, and action in male-dominated environments. *Organizational Behavior and Human Decision Processes* **142**, 28–44 (2017).

15. S. L. Bem, *The Measurement of Psychological Androgyny1* (1974).

16. A. H. Eagly, V. J. Steffen, Gender stereotypes stem from the distribution of women and men into social roles. *Journal of Personality and Social Psychology* **46**, 735–754 (1984).

17. S. T. Fiske, A. J. C. Cuddy, P. Glick, J. Xu, A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology* **82**, 878–902 (2002).

18. A. C. Vial, J. L. Napier, Unnecessary Frills: Communality as a Nice (But Expendable) Trait in Leaders. *Front. Psychol.* **9** (2018).

19. A. S. Rosette, L. P. Tost, Agentic women and communal leadership: How role prescriptions confer advantage to top women leaders. *Journal of Applied Psychology* **95**, 221–235 (2010).

20. M. Infanger, J. Bosak, S. Sczesny, Communality sells: The impact of perceivers' sexism on the evaluation of women's portrayals in advertisements. *European Journal of Social Psychology* **42**, 219–226 (2012).

21. M. Conway, G. Alfonsi, D. Pushkar, C. Giannopoulos, Rumination on Sadness and Dimensions of Communality and Agency: Comparing White and Visible Minority Individuals in a Canadian Context. *Sex Roles* **58**, 738–749 (2008).

22. L. R. James, R. G. Demaree, G. Wolf, rwg: An assessment of within-group interrater agreement. *Journal of Applied Psychology* **78**, 306–309 (1993).

23. T. A. O'Neill, An Overview of Interrater Agreement on Likert Scales for Researchers and Practitioners. *Front Psychol* **8**, 777 (2017).

24. P. D. Bliese, *multilevel: Multilevel functions* (2016).

25. J. W. Pennebaker, R. L. Boyd, K. Jordan, K. Blackburn, The Development and Psychometric Properties of LIWC2015 (2015) (April 4, 2019).

26. R. Rehurek, P. Sojka, *Gensim–python framework for vector space modelling* (NLP Centre, Faculty of Informatics, Masaryk University, 2011).

27. A. E. Abele, *et al.*, Facets of the Fundamental Content Dimensions: Agency with Competence and Assertiveness—Communion with Warmth and Morality. *Front. Psychol.* **7** (2016).

28. A. S. Rosette, C. Z. Koval, A. Ma, R. Livingston, Race matters for women leaders: Intersectional effects on agentic deficiencies and penalties. *The Leadership Quarterly* **27**, 429–445 (2016).

29. M. Sigmund, R. Ferstl, Panel vector autoregression in R with the package panelvar. *The Quarterly Review of Economics and Finance* (2019) https:/doi.org/10.1016/j.qref.2019.01.001 (April 19, 2021).

30. D. Holtz-Eakin, W. Newey, H. S. Rosen, Estimating Vector Autoregressions with Panel Data. *Econometrica* **56**, 1371–1395 (1988).

31. L. P. Hansen, Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica* **50**, 1029–1054 (1982).

32. S. Nickell, Biases in Dynamic Models with Fixed Effects. *Econometrica* **49**, 1417–1426 (1981).

33. M. Arellano, O. Bover, Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics* **68**, 29–51 (1995).

34. R. F. Phillips, A Comparison of First-Difference and Forward Orthogonal Deviations GMM. *arXiv:1907.12880 [econ]* (2019) (April 19, 2021).

35. , How to do Xtabond2: An Introduction to Difference and System GMM in Stata - David Roodman, 2009 (April 19, 2021).

36. J. Alvarez, M. Arellano, The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators. *Econometrica* **71**, 1121–1159 (2003).

37. R. Blundell, S. Bond, Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* **87**, 115–143 (1998).

38. D. W. K. Andrews, B. Lu, Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *Journal of Econometrics* **101**, 123–164 (2001).

39. J. Lin, Teaching notes on impulse response function and structural VAR. *Technical report, Department of Economics, National Chengchi University* (2006).

40. G. Kapetanios, A bootstrap procedure for panel data sets with many cross-sectional units. *The Econometrics Journal* **11**, 377–395 (2008).