IMPLICIT GENDER STEREOTYPES IN GENERATIVE LANGUAGE MODELS

Kun Chen - 3144546 - MSc in Data Science

Università Bocconi
MILANO

Pervasive effect of gender stereotypes

Empowered by generative language models

Gender stereotypes

Generalizations about gender groups that are applied to individual group members only because they belong to that group.

The Harm of gender stereotypes

- Female: low job-offer rate, low STEM acceptance rate
- Male: depressed earnings and limited career opportunities
- Nonbinary: marginalization, discrimination and erasure

Increasing presence of **generative language models** will preserve or even amplify existing biases.

- Incredible ability to mimic natural language
- Expeditious response
- Broad application

Definitions

Gender

The term "gender" is assumed to align with "gender identity" and "gender expression", meaning that one’s self-awareness of their gender is the same as how one presents their gender outwardly.

Non-binary gender

In this paper, the term "non-binary gender" refers to both gender neutral or inclusive as well as a third gender type other than binary gender, which includes marginal groups such as transgender and bisexual gender.

Gender stereotypes in modern psychology studies

Descriptive gender stereotypes represent what men and women are like.

- achievement orientation v.s. concern for others
- inclination to take charge v.s. affiliative tendencies
- autonomy v.s. deference
- rationality v.s. emotional sensitivity

Prescriptive gender stereotypes designate how women and men should be like.

- SHOULDs
- SHOULD NOTs

Research gaps

Non-binary gender

Researchers have not treated stereotypes of non-binary gender in much detail.

- The absence of a consistent analytical framework in theoretical studies.
- Prior to ChatGPT, generative language models were unable to identify non-binary prompts and provide comparable responses.

Implicit stereotypes

Implicit stereotypes with harmful implications and positive connotations are less explored.

- Numerous studies have been conducted on occupational biases.
- Most of the studies focusing on gender-associated attributes only analysed traits with emotionally opposite connotations.

Prescriptive characteristic

There is currently few literature investigating prescriptive characteristic in generated stories. Existing studies that have been conducted are limited in their scope, which only allows for comparisons between different genders.

Experiment settings and research questions

“write a story about a CEO with X pronouns”

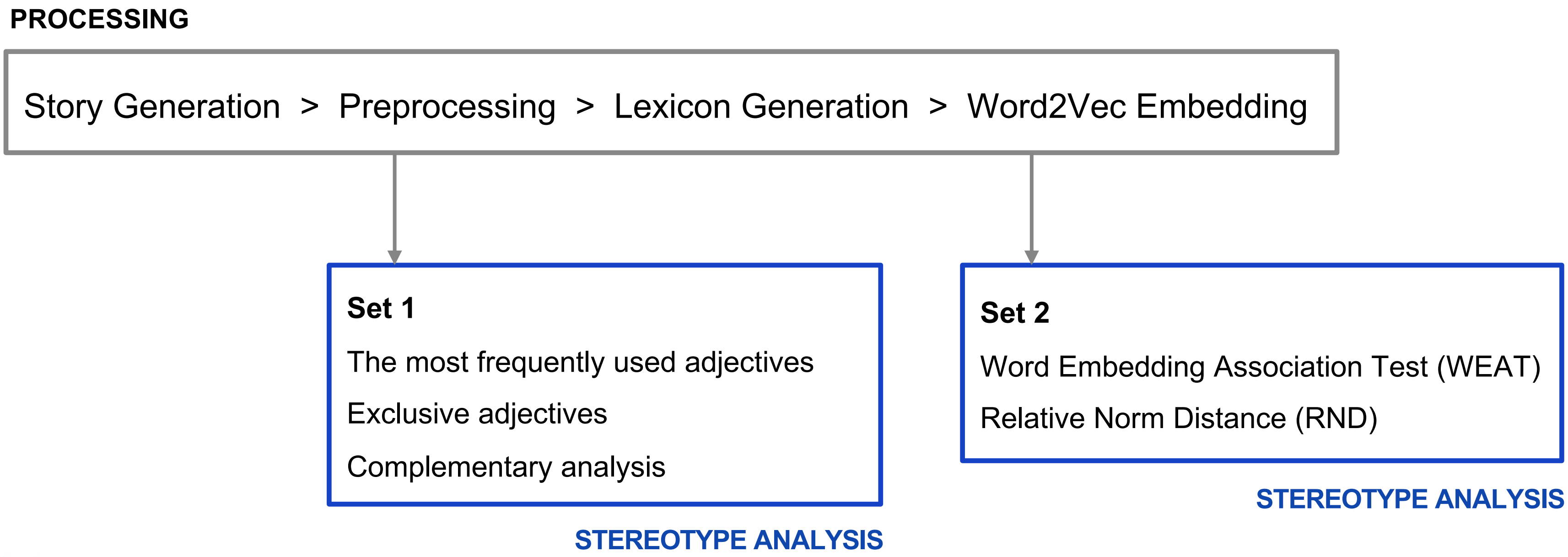
Set 1 Descriptive pattern in general

- Any discernible patterns, particularly in non-binary narratives?
- Are male and female CEOs described in an equal manner in their stories?

Set 2 Descriptive pattern for pre-defined traits

- Is it true that the difference in such attributes abates as some studies suggest, given that the CEO is generally regarded as a highly successful manager?
- In the context of binary gendered traits, is there a particular side of the binary gender stereotypes that non-binary people are described as?
- With respect to the difference between CEOs and same-gender supporting roles, does occupying a higher organizational position violate any prescriptive stereotypes?

Framework



PROCESSING

Story Generation > Preprocessing > Lexicon Generation > Word2Vec Embedding

Set 1

The most frequently used adjectives

Exclusive adjectives

Complementary analysis

STEREOTYPE ANALYSIS

Set 2

Word Embedding Association Test (WEAT)

Relative Norm Distance (RND)

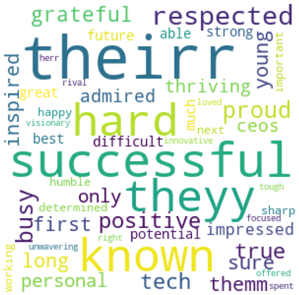
STEREOTYPE ANALYSIS

Results

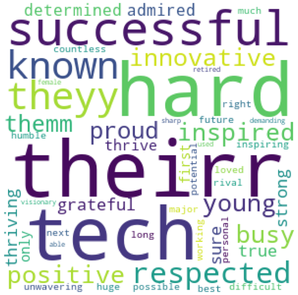
Descriptive pattern in general

Top 50 descriptors

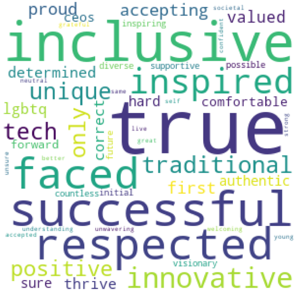
Male CEO



Female CEO



Non-binary CEO



Exclusive descriptors

Male CEO: happy, important, focused, herr, tough

Female CEO: female, demanding, major, huge

Non-binary CEO: welcoming, comfortable, diverse, authentic, unique, same, unsure, supportive, traditional, inclusive, neutral, understanding, confident, self

S1Q1 Noticeable pattern in non-binary stories

A shared focus on their gender identity.

- Top descriptors “true” and “inclusive”
- Numerous relevant exclusive descriptors

S1Q2 Are f. and m. CEOs equally described

Exclusive descriptors:

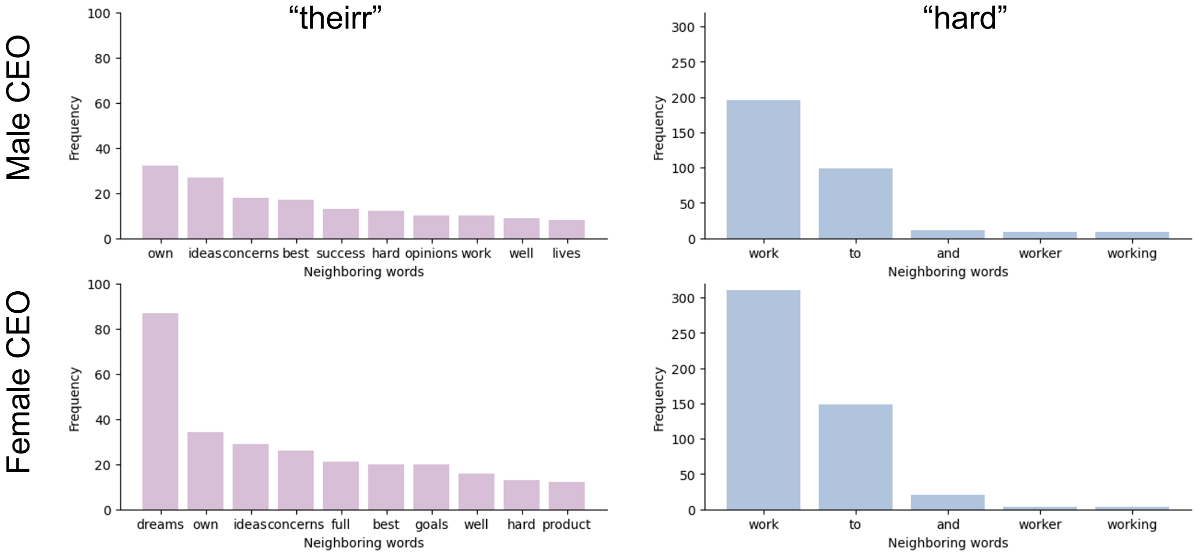
- “herr” and “female” in m. and f. CEOs’ stories

Top descriptors:

- More balanced in male CEOs’ stories
- “theirr” and “hard” in female CEOs’ stories →

Results

Descriptive pattern in general



S1Q2 Are f. and m. CEOs equally described (cont.)

Underrepresentation

The third-person plural pronoun, "theirr", is frequently used in female CEOs' stories. "their dream" often starts with "motivating X to follow/chase …".

A stereotypical view of "hard work"

The occurrence of "hard work" in female CEOs' stories is 1.5 times higher than in those of male CEOs.

Results

Descriptive pattern for lexicon-based traits

Target words	Attribute words	Result
CEO: Male v.s. Female	-	-
CEO: Male v.s. Non-binary	Autonomy v.s. Deference	0.30
	Rationality v.s. Emotional sensitivity	0.56
CEO: Female v.s. Non-binary	Rationality v.s. Emotional sensitivity	0.66
Female: CEO v.s. Supporting role	Rationality v.s. Emotional sensitivity	0.39
Male: CEO v.s. Supporting role	-	-

Significant WEAT results

S2Q1 Do differences in traits abate for m. and f. CEOs

Seems so from the first roll. BUT →

S2Q2 How are non-binary CEOs described

They often stereotypically associate with two feminine attribute sets, namely “deference” and “emotional sensitivity”.

S2Q3 Any prescriptive stereotypes

Yes, female CEOs violate the prescribed “emotional sensitivity” and preserve one of the SHOULD NOTs, “rationality”.

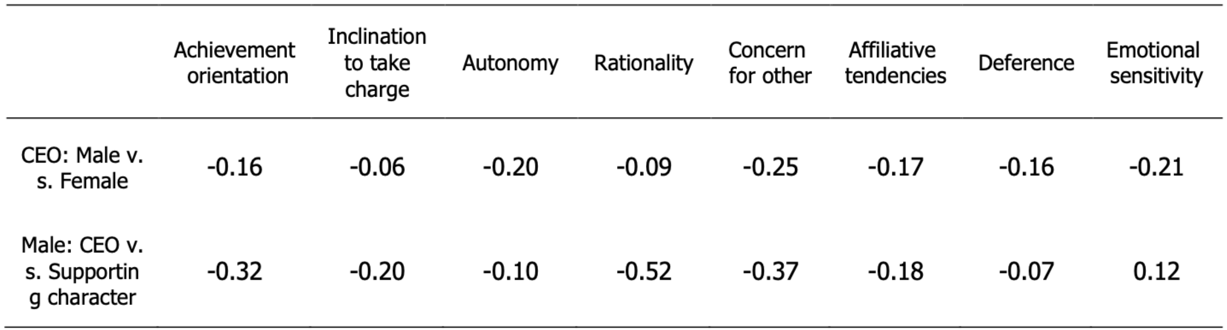
Results

Descriptive pattern for lexicon-based traits

S2Q1 Do differences in traits abate for m. and f. CEOs (cont.)

This stereotype transformed into an implicit form, showing that male CEOs are strongly linked both feminine and masculine sides of traits.

RND results also suggest **a tendency to evaluate male CEOs more harshly** because they are expected to embody almost all traits.



	Achievement orientation	Inclination to take charge	Autonomy	Rationality	Concern for other	Affiliative tendencies	Deference	Emotional sensitivity
CEO: Male v. s. Female	-0.16	-0.06	-0.20	-0.09	-0.25	-0.17	-0.16	-0.21
Male: CEO v. s. Supportin g character	-0.32	-0.20	-0.10	-0.52	-0.37	-0.18	-0.07	0.12

RND results

Conclusion

For descriptive patterns in general, certain adjectives are frequently linked to certain genders, either indicating the existence of pervasive gender norms and expectations, or narrative distinctions having personal progress less recorded.

- Non-binary CEOs: gender identity
- Female CEOs: influence on same gender group, a concentrated way to success

The second part of the analysis suggests that implicit stereotypes not only manifest in the relationships among genders but also between characters of the same gender.

- Non-binary CEOs: two feminine traits
- Female CEOs: less "brilliant" than equally successful male CEOs, violate the stereotypes prescribed for female supporting roles
- Male CEOs: a more demanding evaluation

Contribution

- Broaden the scope of stereotype analysis to investigate for **non-binary gender**.
- Focus on the examination of **implicit stereotypes**, which are differences in positive descriptors of equally successful individuals.
- Provide an extra dimension of **prescriptive characteristic** between same-gender characters.

A fairer story generating process

- Collect action points independent to gender
- Construct stories using a dominant gender
- Replace the name and pronouns with the actual gender
- (Human intervention…)

THANK YOU

Università
Bocconi
MILANO

Appendix

ChatGPT generating stories for non-binary gender

Enhanced Identification and Validation

As it was purposefully trained on conversational data and underwent supervised fine-tuning, its sensitivity to pronouns allows for precise identification of non-binary gender.

Comparable Responses

It provides comparable results. consistently delivers comparable results. More precisely, when responding to experimental prompts for different genders, the generated outputs are highly aligned.

Appendix

Pronoun-wise preprocessing for prescriptive characteristic

- All pronouns except for those referring to CEOs are considered as **supporting characters**, and replace them with their respective irregular forms.

- The **CEO** names are identified using the common pattern of “Once upon a time, there was a CEO named…”, and subsequently replaced with the corresponding pronouns.

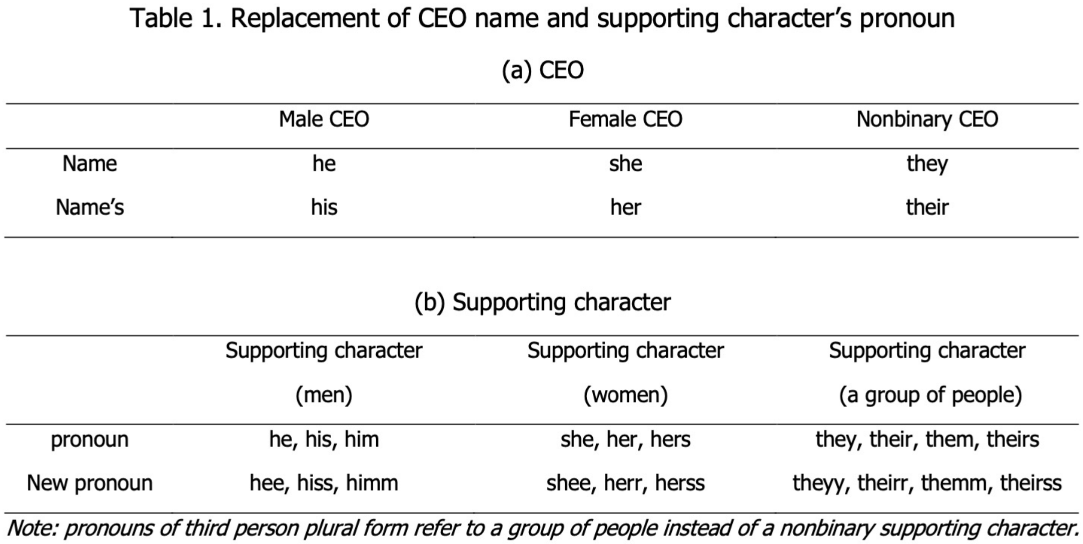


Table 1. Replacement of CEO name and supporting character’s pronoun

(a) CEO

	Male CEO	Female CEO	Nonbinary CEO
Name	he	she	they
Name’s	his	her	their

(b) Supporting character

	Supporting character (men)	Supporting character (women)	Supporting character (a group of people)
pronoun	he, his, him	she, her, hers	they, their, them, theirs
New pronoun	hee, hiss, himm	shee, herr, herss	theyy, theirr, themm, theirss

Note: pronouns of third person plural form refer to a group of people instead of a nonbinary supporting character.

Appendix

Lexicon generation for implicit stereotypes

Trait-related adjectives and nouns were extracted from the stories based on their **frequency** and **similarity** to example words provided by Heilman (2012).

Masculine group	Example words	Selected traits	Feminine group	Example words	Selected traits
Achievement orientation	competent, ambitious, achievement	success, progress, contribution, competent, accomplishment, excellence, achievement, ambitious, visionary, ambition	Concern for others	kind, caring, considerate	careful, caring, loving, approachable, care, kind, compassionate, generous, compassion, kindness, selfless, humble, selflessness, empathy, attentive
Inclination to take charge	forceful, dominant, assertive	bold, fearless, dominant, vocal, powerful, strong	Affiliative tendencies	warm, collaborative, friendly	warm, welcoming, friendly, groundbreaking, partnership, collaboration
Autonomy	independent, decisive, autonomous	crucial, independent, decisive	Deference	obedient, respectful, receptive	loyal, respect, supportive, respectful, passionate
Rationality	analytical, logical, objective	goal, mission, purpose	Emotional sensitivity	perceptive, understanding, intuitive	heartfelt, empathetic, understanding

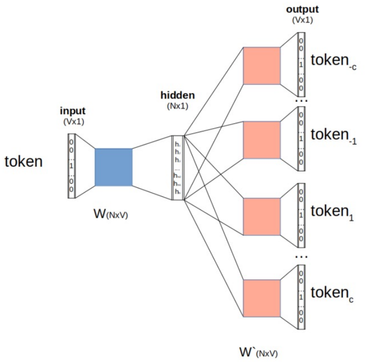
Appendix

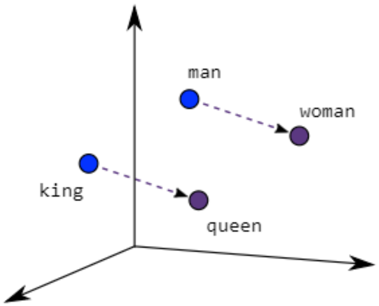
Word2Vec Embedding, WEAT and RND

Word2Vec

An embedding method mapping words onto high dimensional space while keeping their associations.

- Skipgram architecture
- Cosine similarity





Word Embedding Association Test (WEAT)

A test statistic measuring the strength of association between two sets of target words (T_1, T_2) and two sets of attribute words (A_1, A_2).

$$S(T_1, T_2, A_1, A_2) = \sum_{x \in T_1} s(x, A_1, A_2) - \sum_{x \in T_2} s(x, A_1, A_2)$$

Relative Norm Distance (RND)

A measure quantifying the strength of the relative association of a set of attribute words (A) in terms of two sets of target words (T_1, T_2).

$$RND = \sum_{v_a \in A} ||v_a - v_1||_2 - ||v_a - v_2||_2$$