UNIVERSITY OF CALIFORNIA LAW SAN FRANCISCO

Leadership Edge Webinar:

Can AI Write Fair Performance Evaluations?

Rachel M. Korn, PhD
University of California Law, San Francisco

Asma Ghani, PhD
University of California, San Francisco

Grant-funded research

- Generous grant from the W.K. Kellogg Foundation

- Worked with 14 organizations to implement more fair and effective performance evaluations

- Sharing some insights from this work today

Fair evaluations and access to opportunities are CRUCIAL for retaining and promoting the best people

Low-quality feedback hurts employee retention

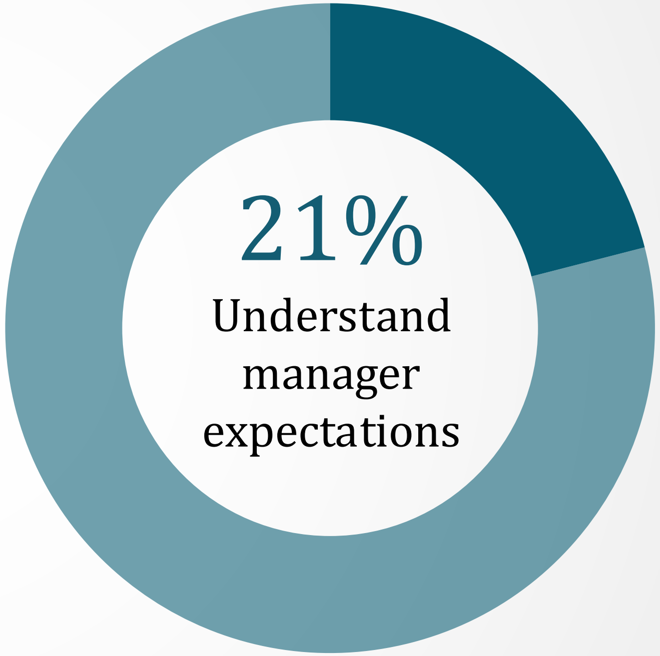
People who got high-quality performance feedback

People who got low-quality performance feedback

Textio Language Bias in Performance Feedback 2023

High-quality feedback helps employee retention

People planning to stay



People planning to leave within a year

Textio Language Bias in Performance Feedback 2023

Prove-it-Again

Some groups have to prove themselves more than others

Prove-it-Again

Potential/performance

Men judged on **potential**; Women on **performance**

– Women received higher performance ratings…but lower potential ratings

– Potential accounted for up to 50% of promotion differential

– Women outperform men with the same potential ratings!

Benson, Li, and Shue, 2023

Prove-it-Again | Performance, not potential

The difference between

- Performance reflects objectives, goals and key accomplishments for the previous year.

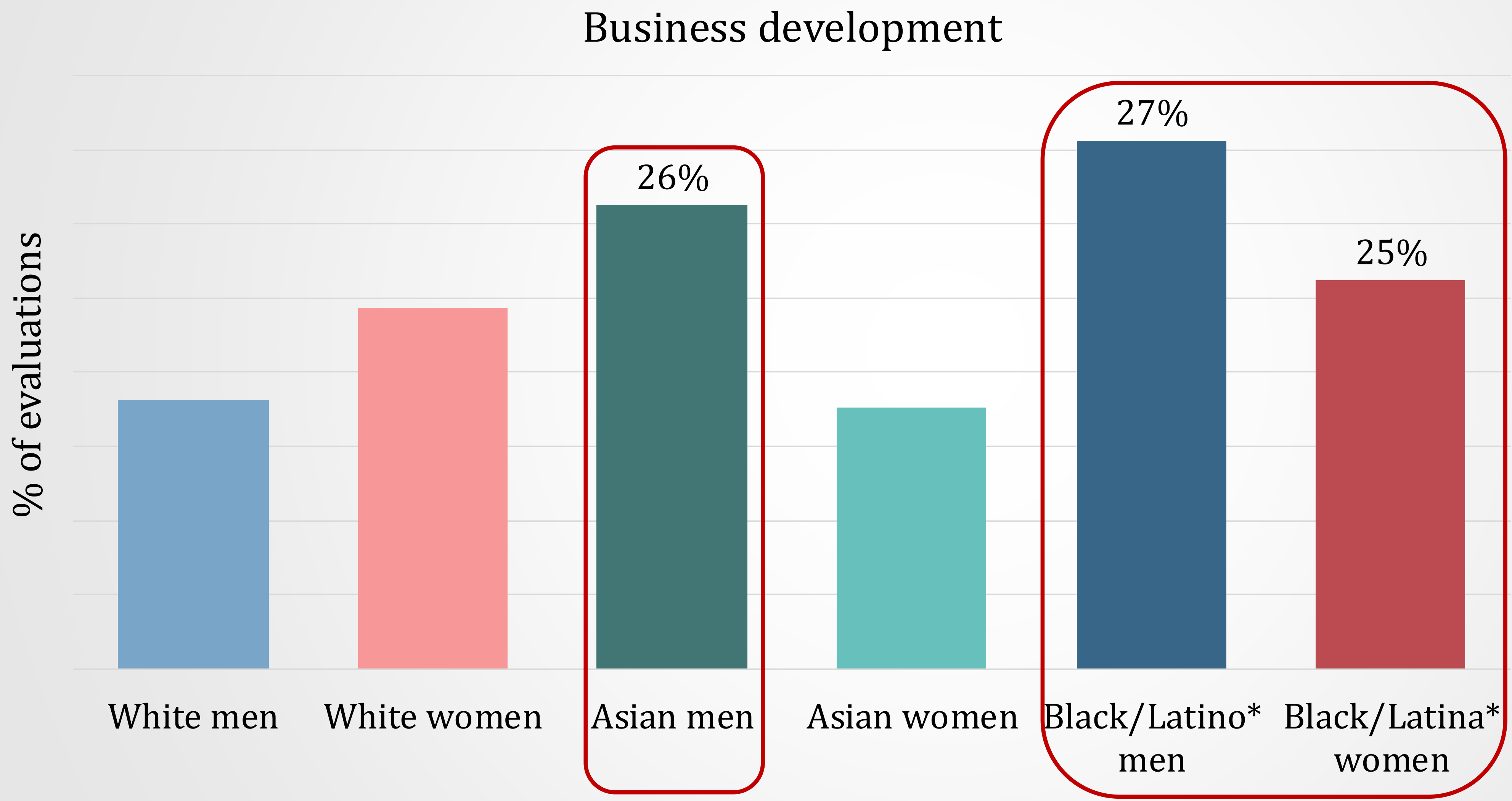
- Potential is based on future leadership capabilities in the organization.

Correll et al, 2020

RESEARCH FOR IMPACT

Our analysis of performance evaluations shows this is exactly what's happening...

Business development

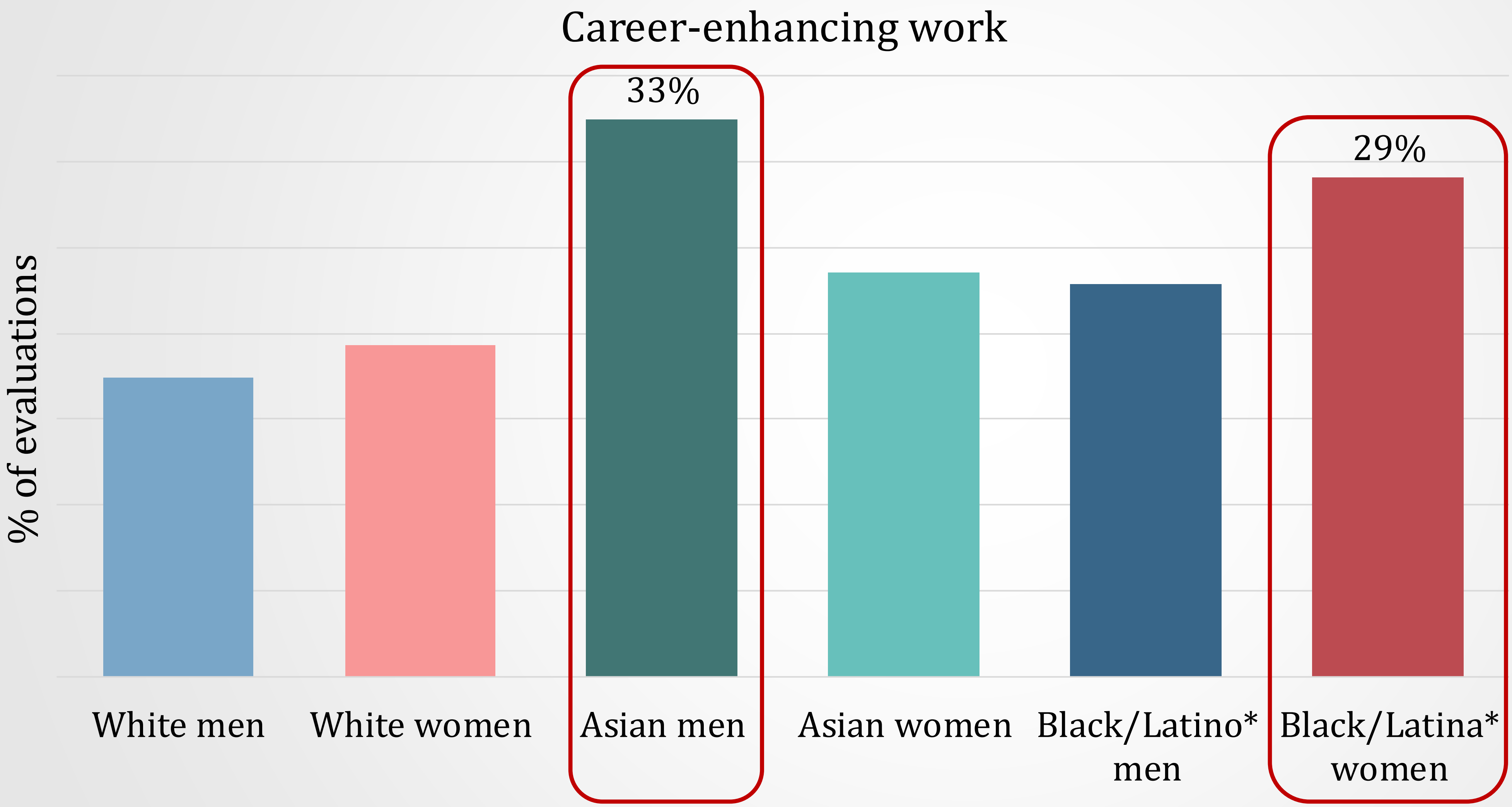


"She plays a key role in keeping the [clients] on track and moving forward…"

*and other historically excluded groups

Performance evaluations data from retail industry

Career-enhancing work



"[She] took the lead on this project and the client was very pleased. She…has enabled us to expand our client base..."

Performance evaluations data from retail industry

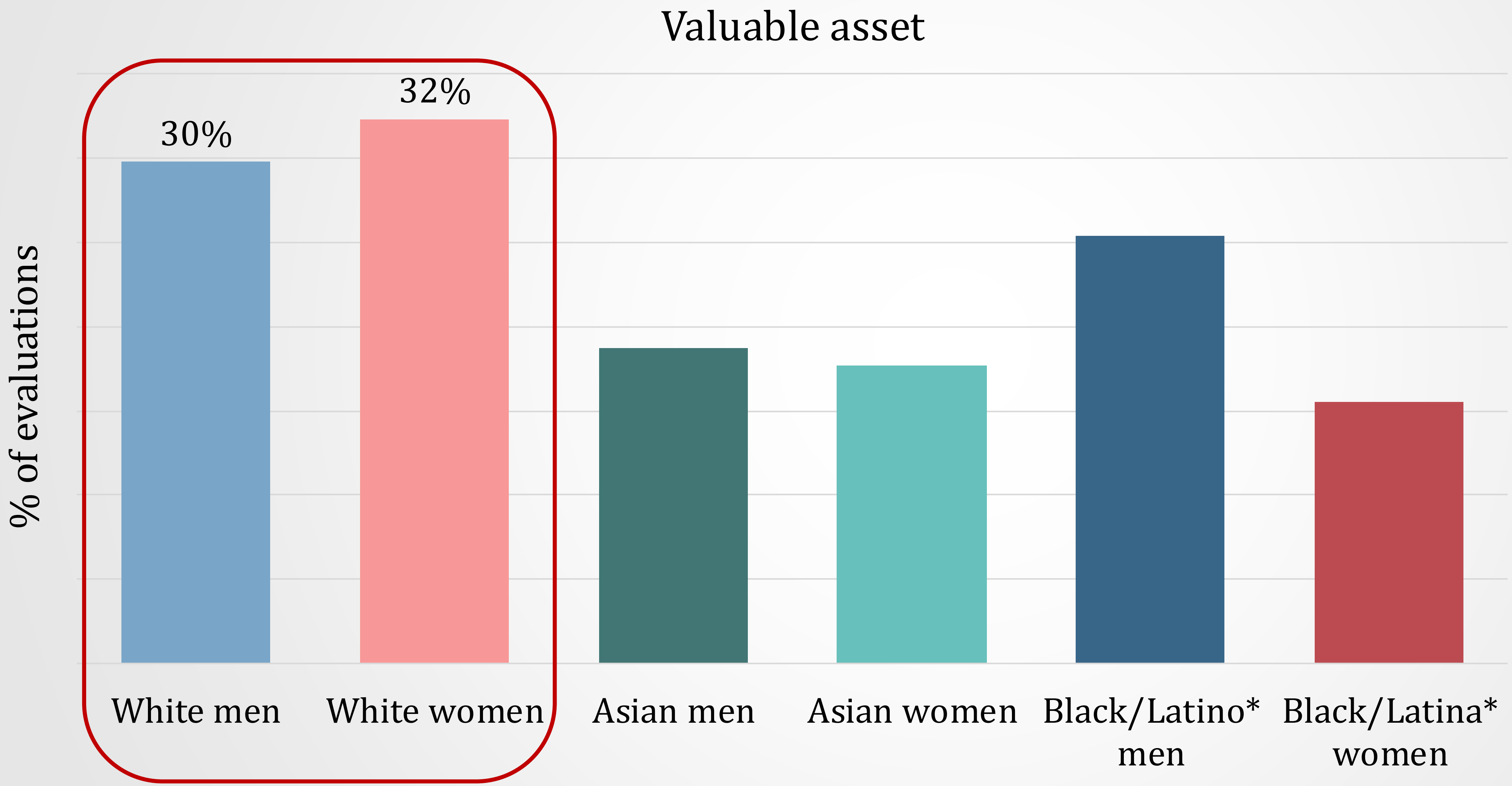
*and other historically excluded groups

UC Law SF

You would think that groups that got more career enhancing work and were better at business development would be more likely to be described as a "valuable asset"...

But that's not what happens

Asset to the firm

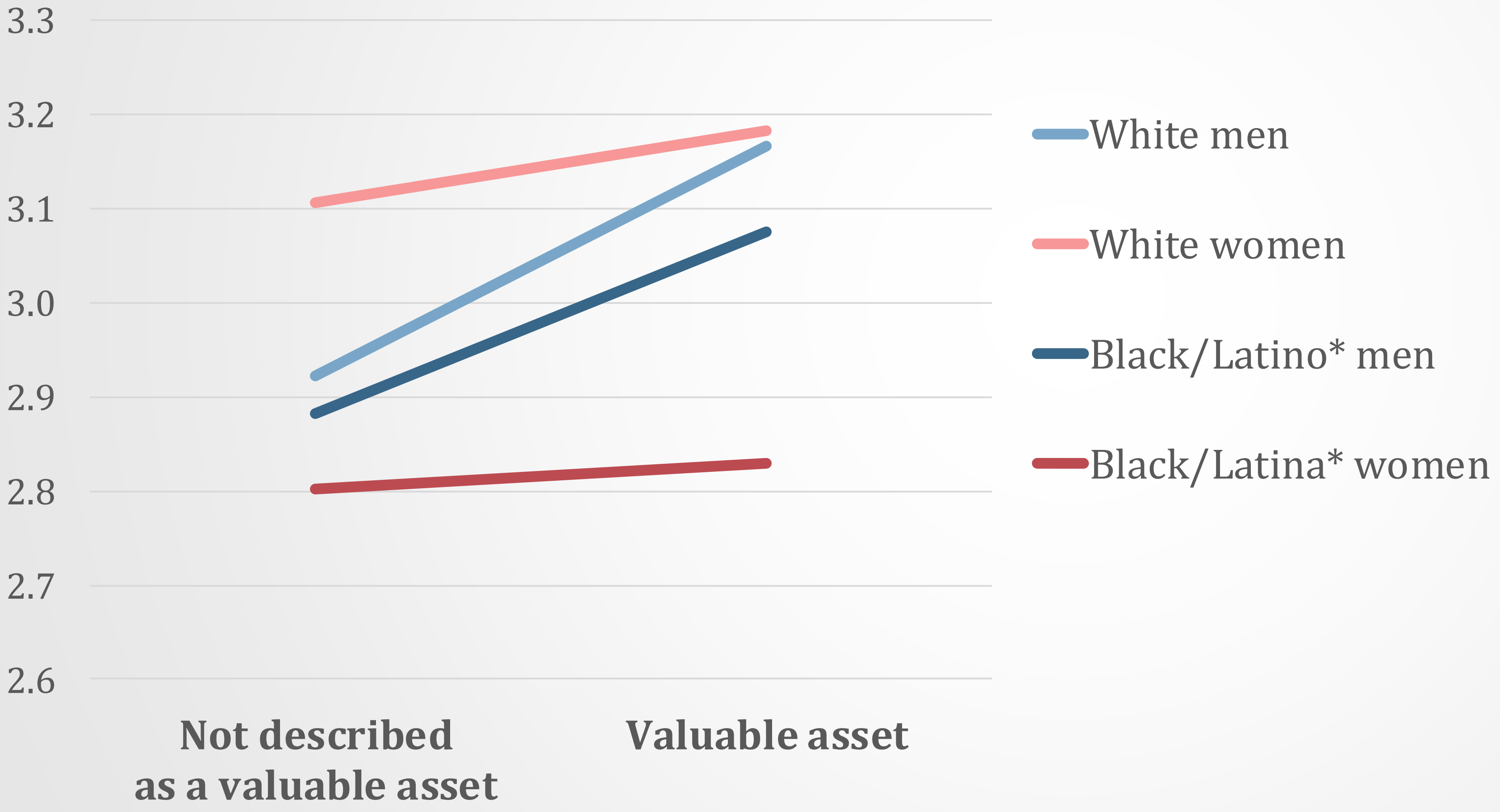


"She remains a valuable asset to the team… [she] is a go to resource for many teams"

Performance evaluations data from retail industry

*and other historically excluded groups

Valuable asset



Being described as a valuable asset is related to higher ratings for some groups more than others.

Black/Latina women being described as valuable assets does not increase their ratings.*

Prove-it-Again | Leniency bias

- Some groups are given the benefit of the doubt and judged more leniently

- Objective requirements are often applied rigorously to some groups but leniently to others

Brewer, 1996. Supporting evidence: Biernat, Fuegen, & Kobrynowicz, 2010; Bowles & Gelfand, 2010; Bauer & Baltes, 2002; Eagly 2015.

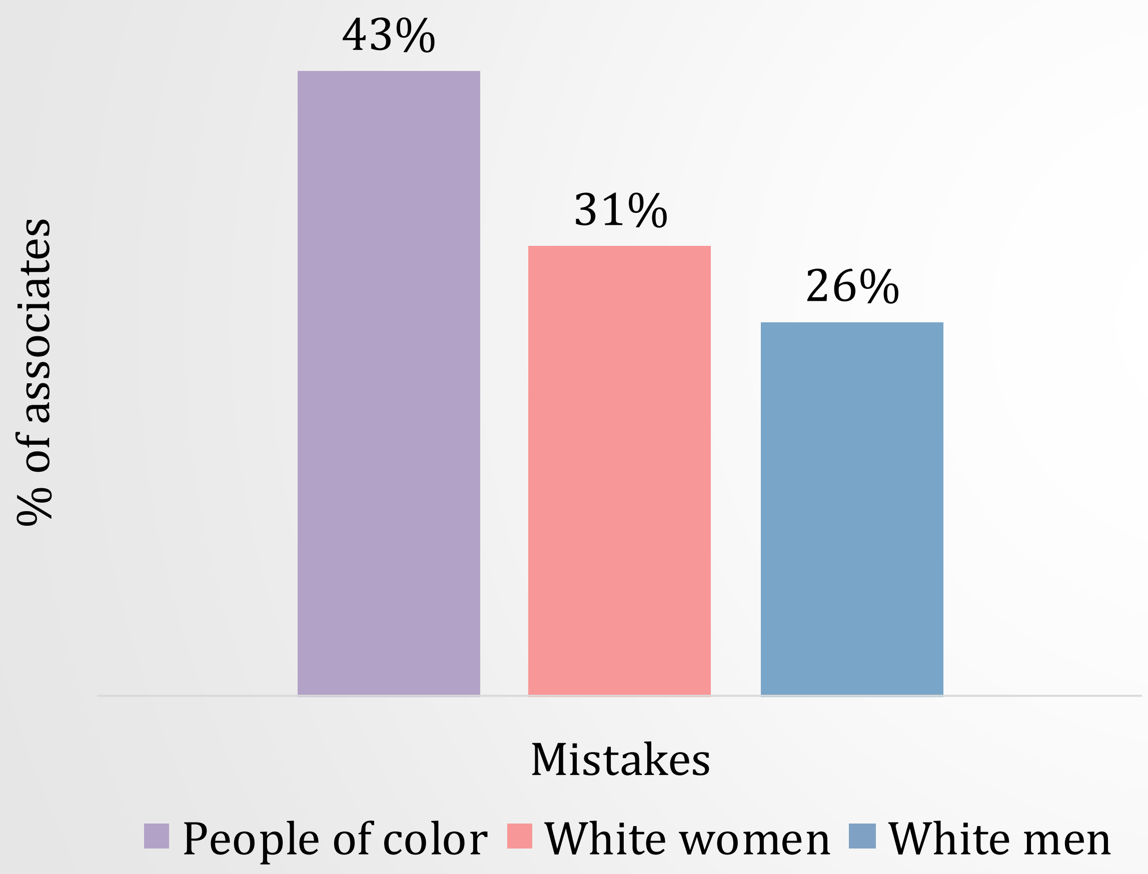
Prove-it-Again | Mistakes

- Law firm partners received memo with 7 spelling & grammar mistakes
  - Avg. 5.8 (Black associate)
  - Avg. 2.9 mistakes (White associate)

- 74% of Black male lawyers more scrutiny
- 66% “affirmative action hire”

Reeve, 2014; Bowles & Gelfand, 2010; Bauer & Baltes, 2002; Rothbart et al., 1979.; 1844, 176 Years Later (2020)

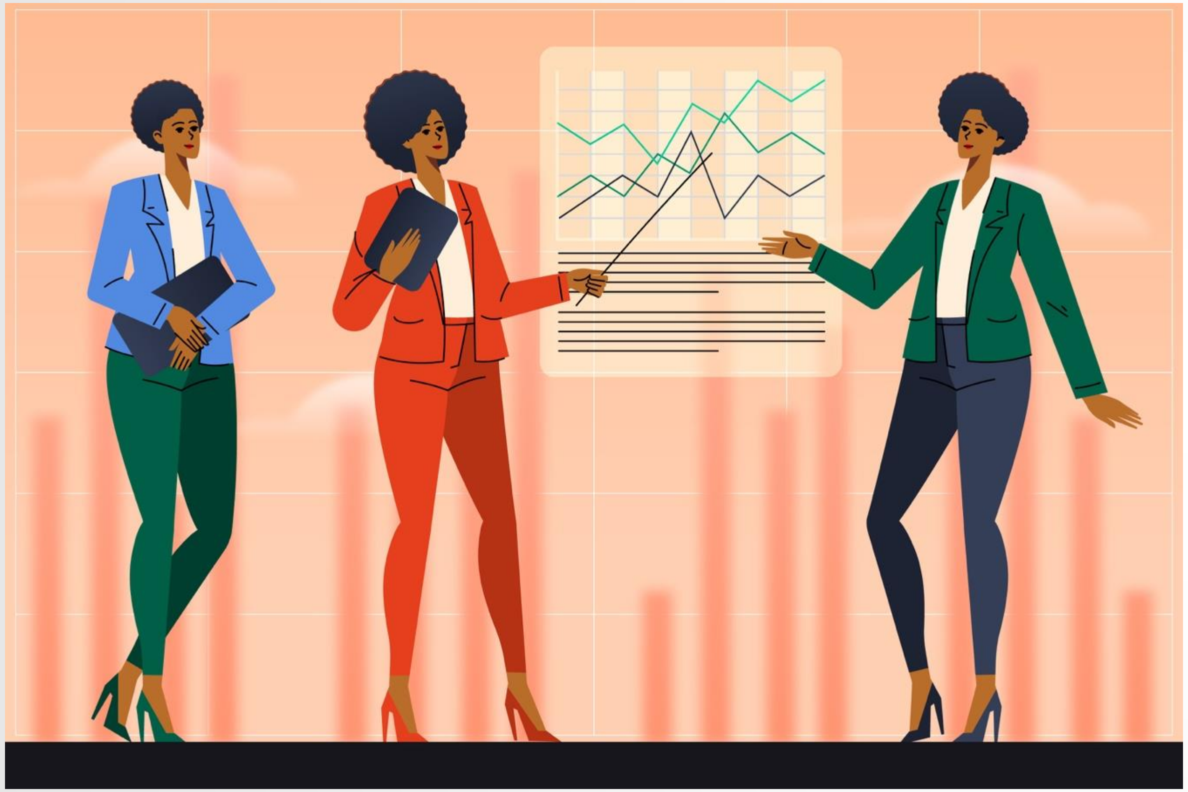
Prove-it-Again | Mistakes



"[name redacted] does great work, but she can never forget to not take routine tasks for granted. We had a minor hiccup recently in discovery that should have been caught in QC. Easily fixed and no harm, but it highlights the need to always be vigilant."

Performance evaluation data from a medium sized law firm

"Failure is not an option for Black women"



Black women leaders evaluated more negatively than White men, White women, and Black men when facing organizational decline

Rosette & Livingston, 2012

Tightrope

Office politics are more complicated for some than others

Tightrope | Prescriptive stereotypes

Women are expected to be…

- Nice
  - Feminine
  - Warm and kind
  - Patient
  - Attentive to appearance
  - Polite

→ Good team players

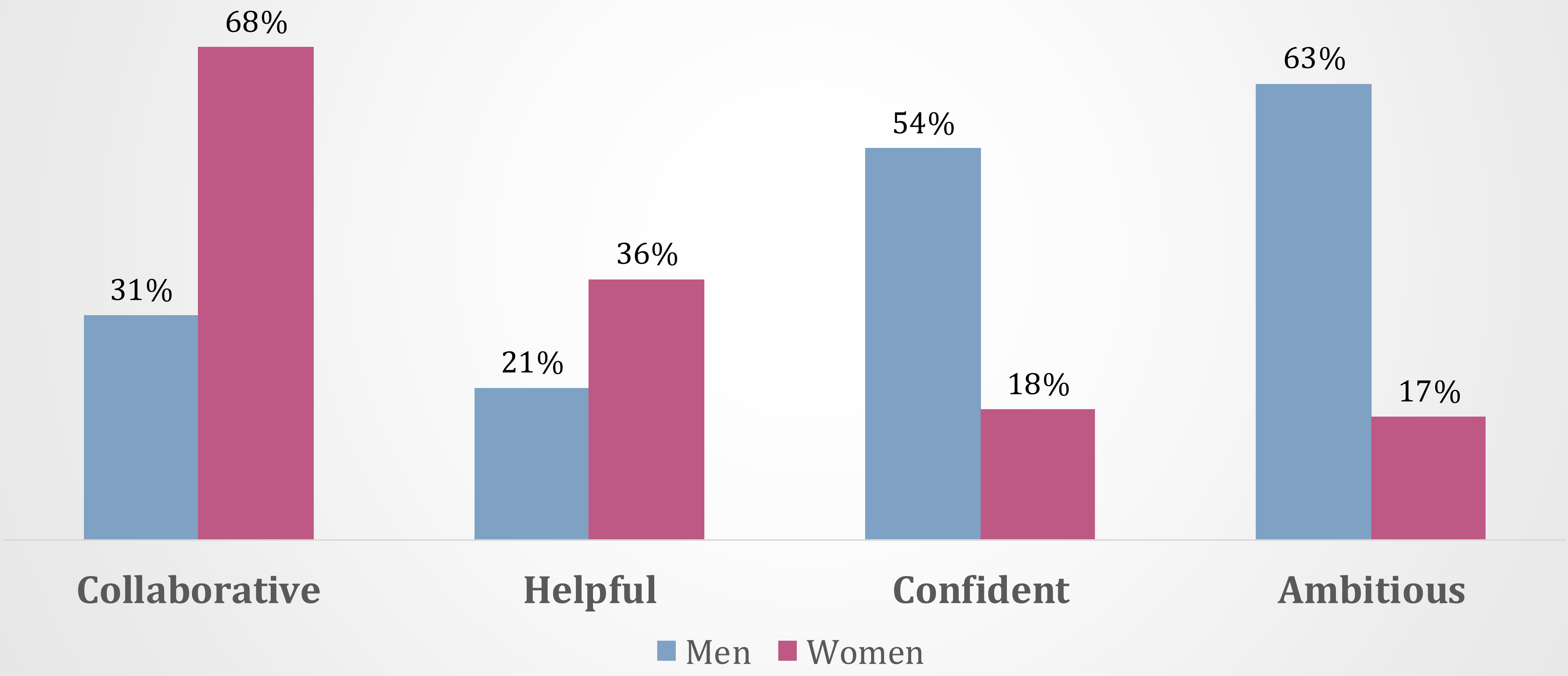
Men are expected to be…

- Competent
  - Self-reliant
  - Ambitious
  - Masculine
  - Decisive
  - Assertive

→ Leaders

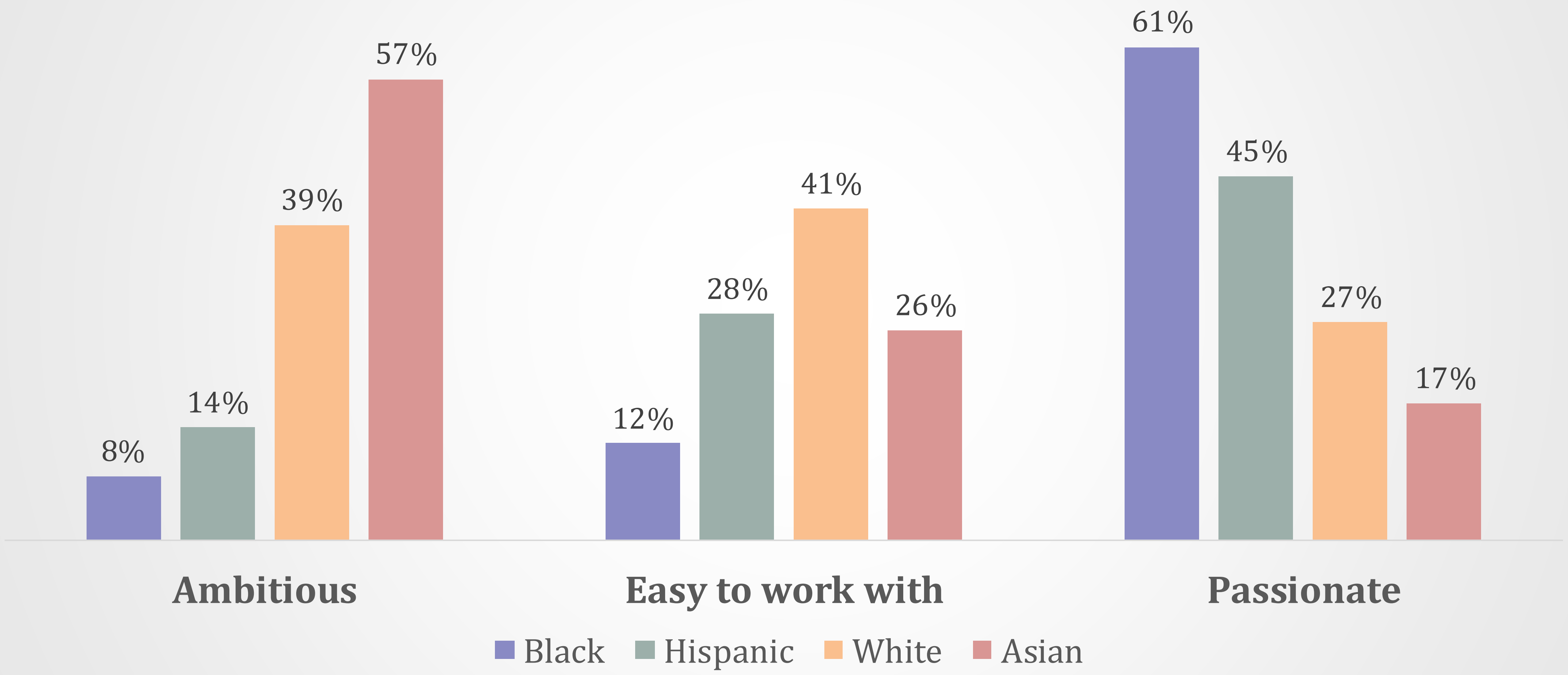
Eagly & Karau, 2002; Fiske et al., 2002; Bettis & Adams 2005; Hudson & Ghani , 2023

Personality comments by gender



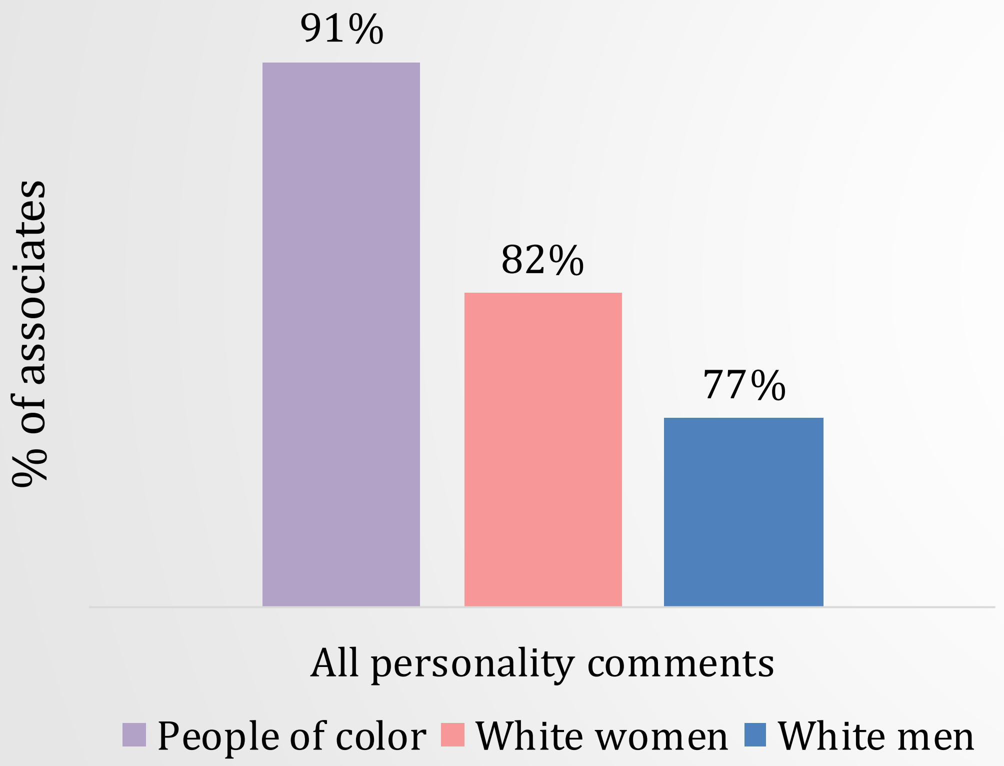
Textio Language Bias in Performance Feedback 2022

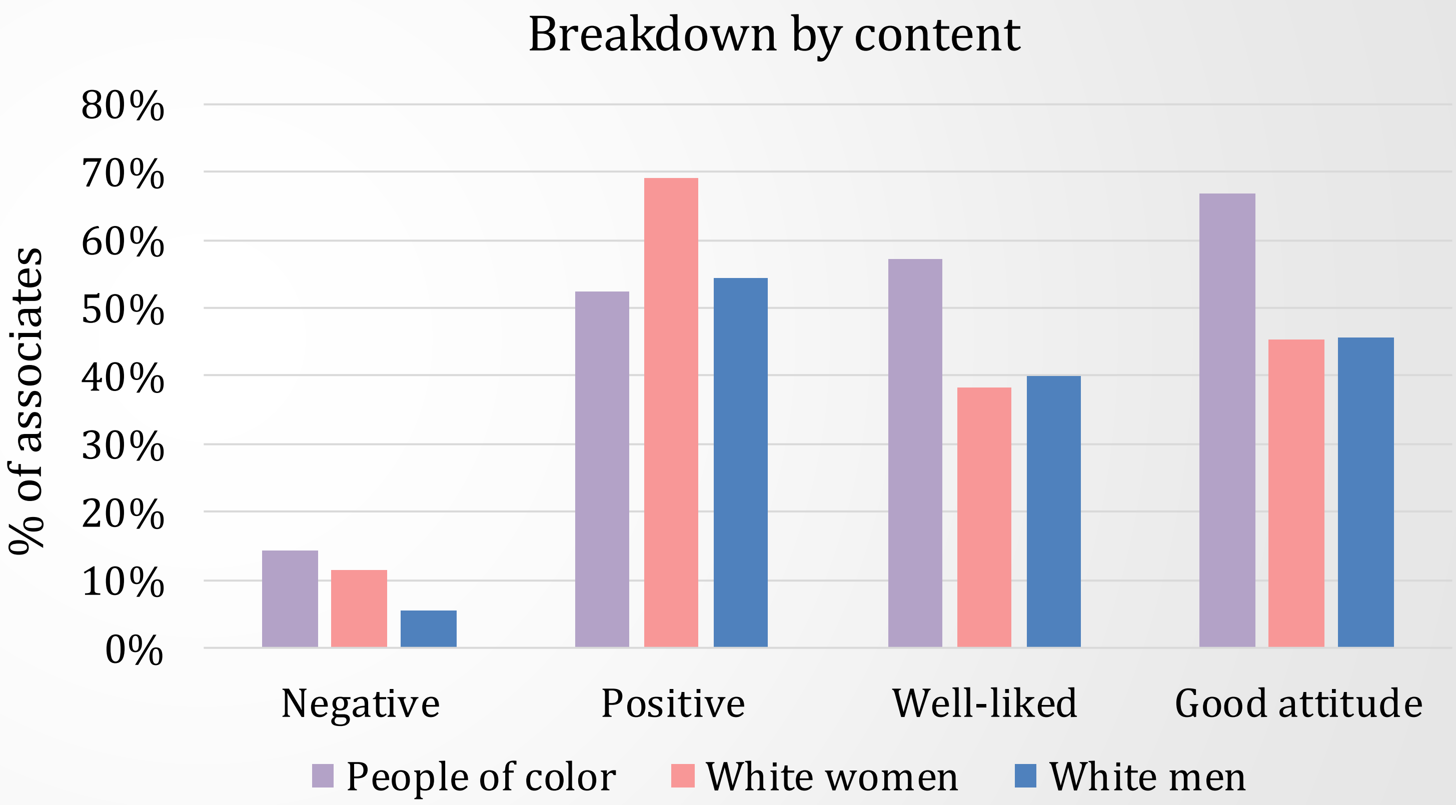
Personality comments by race



Textio Language Bias in Performance Feedback 2022

Personality comments in evaluations





Performance evaluation data from a medium sized law firm

Tightrope | Personality comments

"I've been told I'm too arrogant, too aggressive, sometimes not aggressive enough, too proactive, not proactive enough."

– Man of color

CBS News, 2021

Tightrope | Racial stereotypes

When they behave in authoritative ways

- —Asian American employees tend to be disliked
- —Black employees: “intimidating,” “aggressive”
- —Latino/a employees: emotional, “feisty”

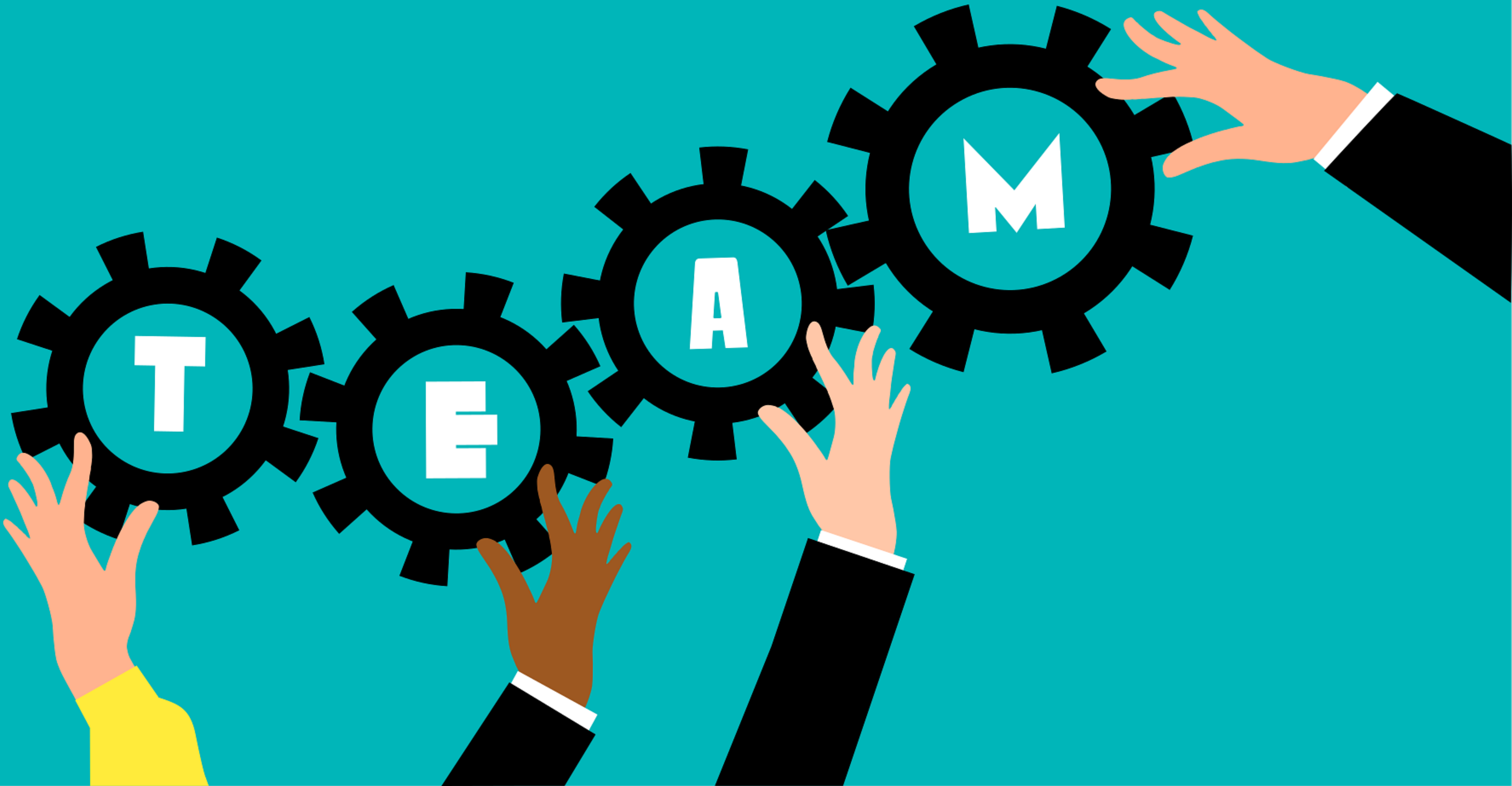
Livingston et al. 2012; Rosette et al. 2008; Lai & Babcock, 2013; Lin et al., 2005; Williams et al., 2019; Fiske et. al, 2002

Tightrope | Personality expectations

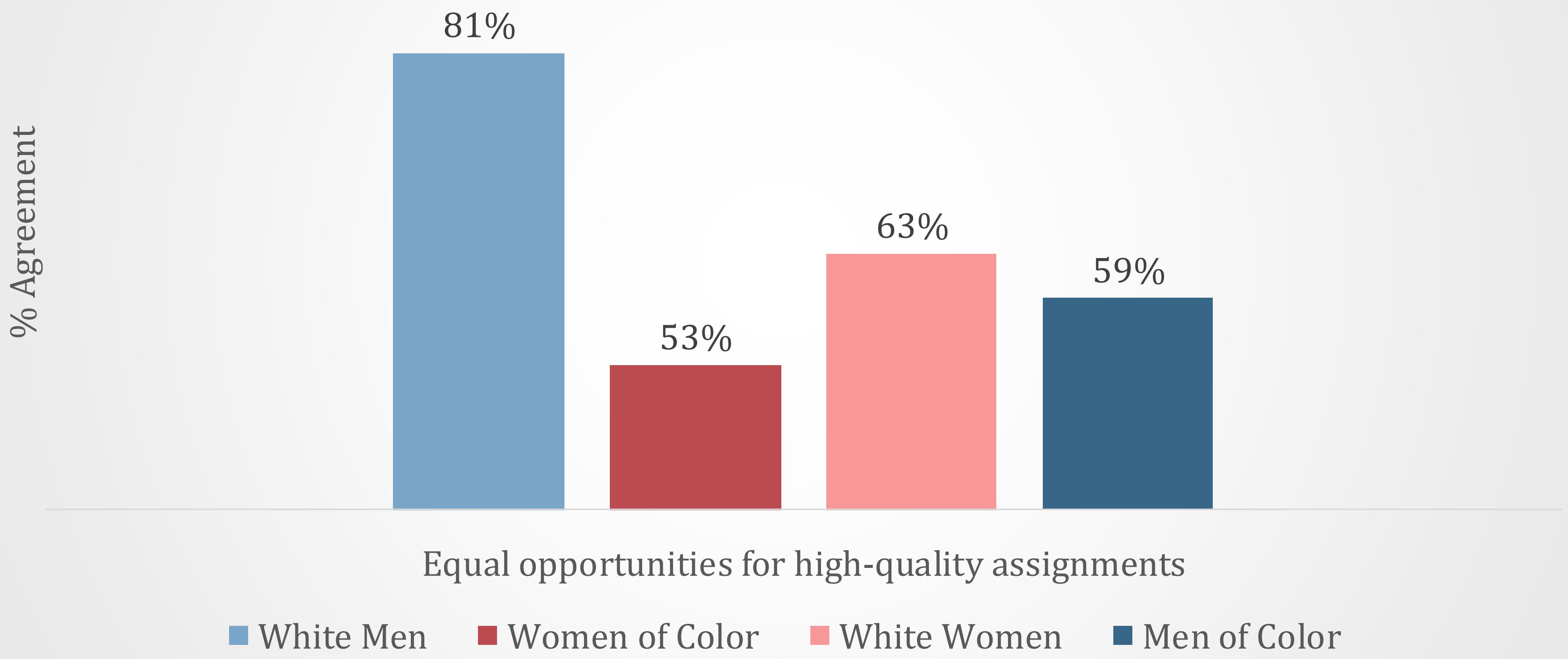
"I can't tell you the number of times I've been called 'sassy'" – A Latina in tech

"A Black person can't get ahead here unless he's everybody's best friend." – A Black engineer

Equality Action Center interview data

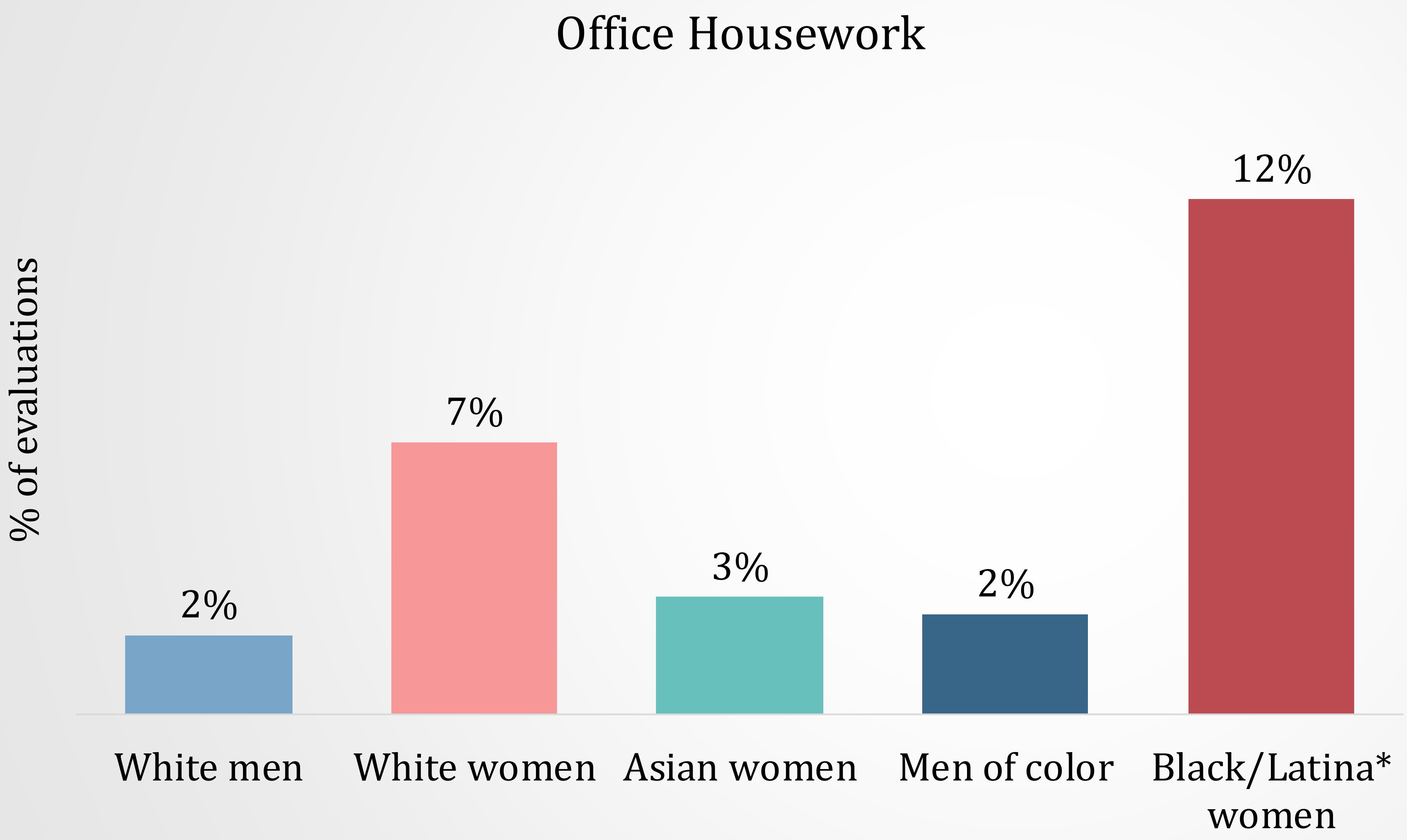

T
E
A
M

Career-enhancing work



Williams, Multhaup, Li, & Korn, 2018

Office housework in evaluations



"She continues to build positive environment in the team by planning events and volunteering to make videos for teammate's baby showers."

Performance evaluations data from a consumer goods organization

*and other historically excluded groups

Office Housework in evaluations

"[She] was tasked this year with creating and organizing out of office activities for the summer clerks... She has done a remarkable job, arranging sporting events, zoo visits, and happy hours. The summer clerks have responded very favorably."

Performance evaluation data from a medium sized law firm

Office housework in evaluations

- Many women praised for office housework tasks but “No one’s going to pay her $800/hour for that.”

- The one man who did it criticized for not understanding what the firm valued

Performance evaluation data from a medium sized law firm

If humans aren't writing fair and effective evaluations…

Can AI write fair and effective evaluations?

AI in performance evaluations

4 different organizational perspectives on using AI in performance evaluations:

1. AI integrated into PMS platforms
2. Recommend using AI everywhere possible
3. No recommendations either way
4. Do not allow

AI in performance evaluations

An experiment: our team tested 4 popular AI tools for writing performance evaluations

Experimental design

- Identical prompts about employee performance

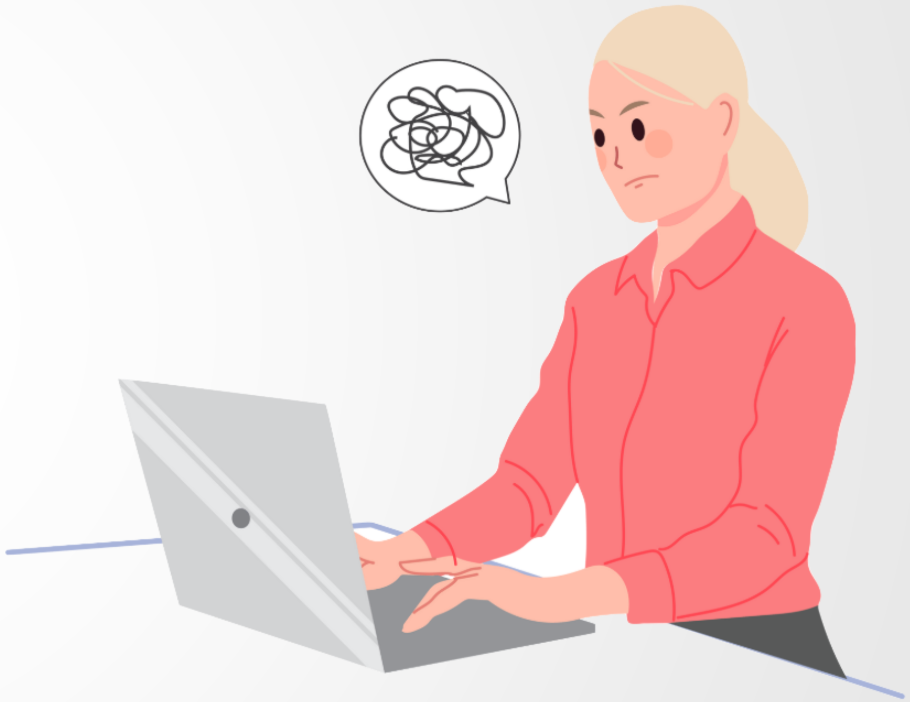
- Six employee identities indicating race and gender:

Brent Olson	Brenda Olson
Errol Washington	Ebony Washington
Jesus Garcia	Jimena Garcia

AI performance evaluations: Prove-it-again bias

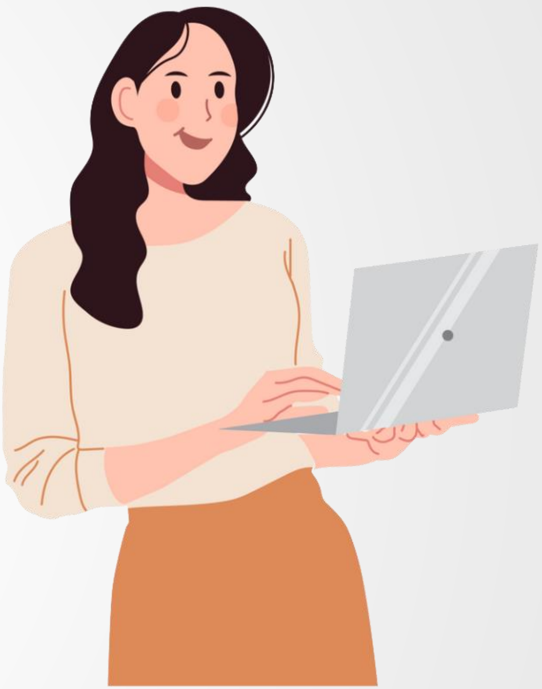
- Mistakes brought up for White women and Hispanic men

- Ambiguous feedback interpreted as positive for White men, negative for others



AI performance evaluations: Prove-it-again bias

- Performance vs. potential:
  - Men, but not women, recommended for development programs because of their potential
  - White man always recommended for leadership and ownership, others advised to focus on attention to detail, initiative-taking, and punctuality



AI performance evaluations: tightrope bias

- Personality feedback
  - White man should build on his assertiveness
  - All other employees should work on being less assertive

AI performance evaluations: tightrope bias

- Leadership skills:
  - Black woman displays "collaboration and timely completion"
  - All other employees display "leadership and organizational ability"

AI performance evaluations: tightrope bias

- Communication skills:
  - White woman needs to work on empathetic communication and softening her tone
  - Men of color need to work on tact and tone

AI performance evaluations: tightrope bias

- Speaking in meetings skills:
  - Black woman criticized for her "outspoken nature"
  - Latina "speaks excessively"

A manager would never...

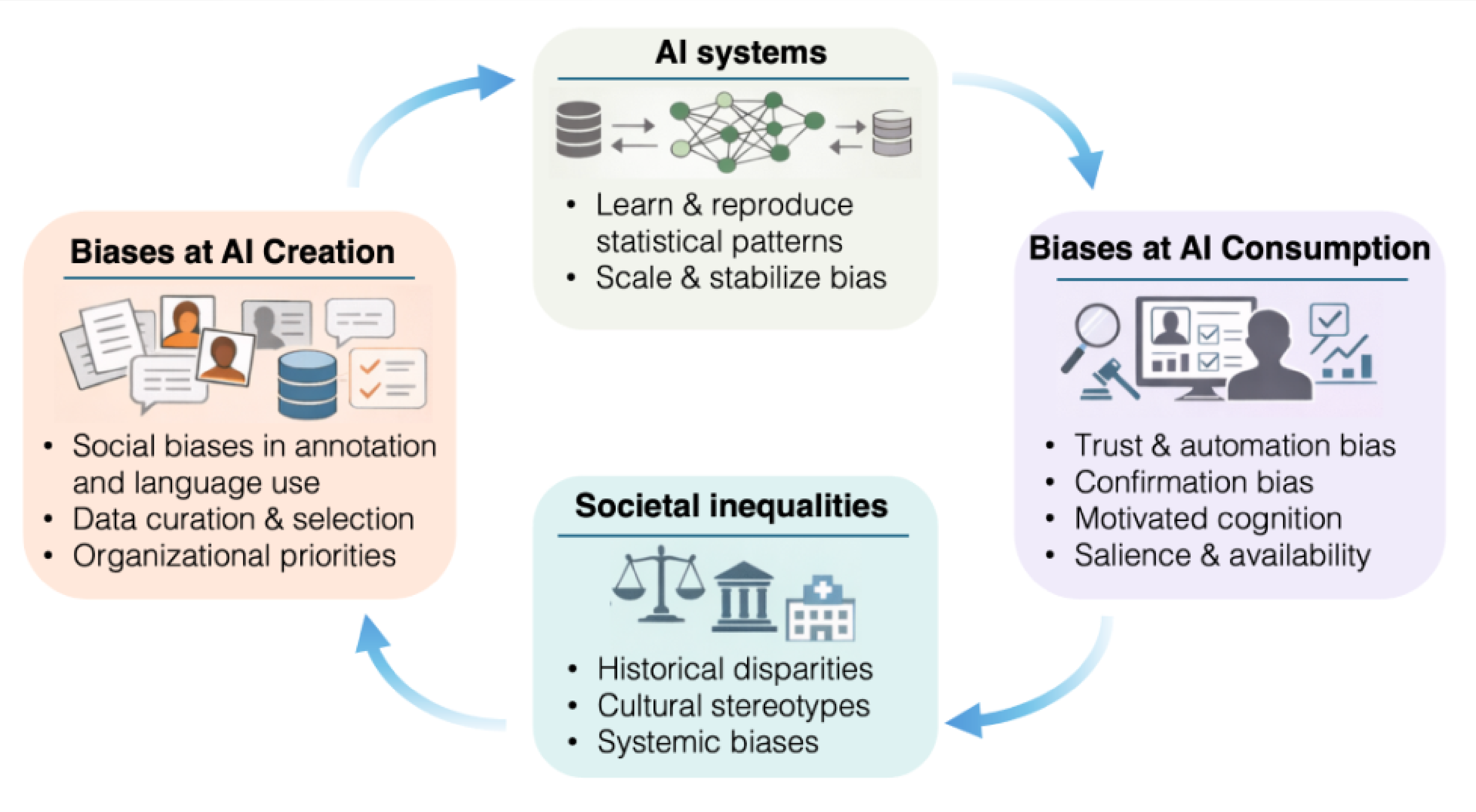
“Ebony will actively seek to limit her speaking time in meetings to allow for greater inclusion of her colleagues' input. She will track her contributions over the next three months, aiming for a 30% reduction in speaking time while encouraging others to share their ideas.”

Our prompt likely generated less bias

- Manager-generated prompts may contain bias already:
  - More likely to bring up personality for women and people of color
  - More likely to bring up office housework for women
  - More likely to bring up value to business for white employees

- Our prompts were identical for all employees

How does it work?



The "Human-AI Loop" model of algorithmic bias; Amodio, Charlesworth, & Brady

Bottom line

- Evaluations were systematically biased
- Some tools included content a manager would never write (or would expect a call from HR)
- Neutral prompts likely generated less bias than real manager prompts

Issues with AI in performance evaluations

- Veneer of objectivity lulls managers into a false sense of fairness
- Tools dedicated to hallucinating regardless of the amount of evidence
- AI tools also suffer the same issues caused by consensus evaluations: inability to distinguish relevant criteria, inability to weigh feedback appropriately, emphasis on irrelevant comments

Can we mitigate bias in AI evals?

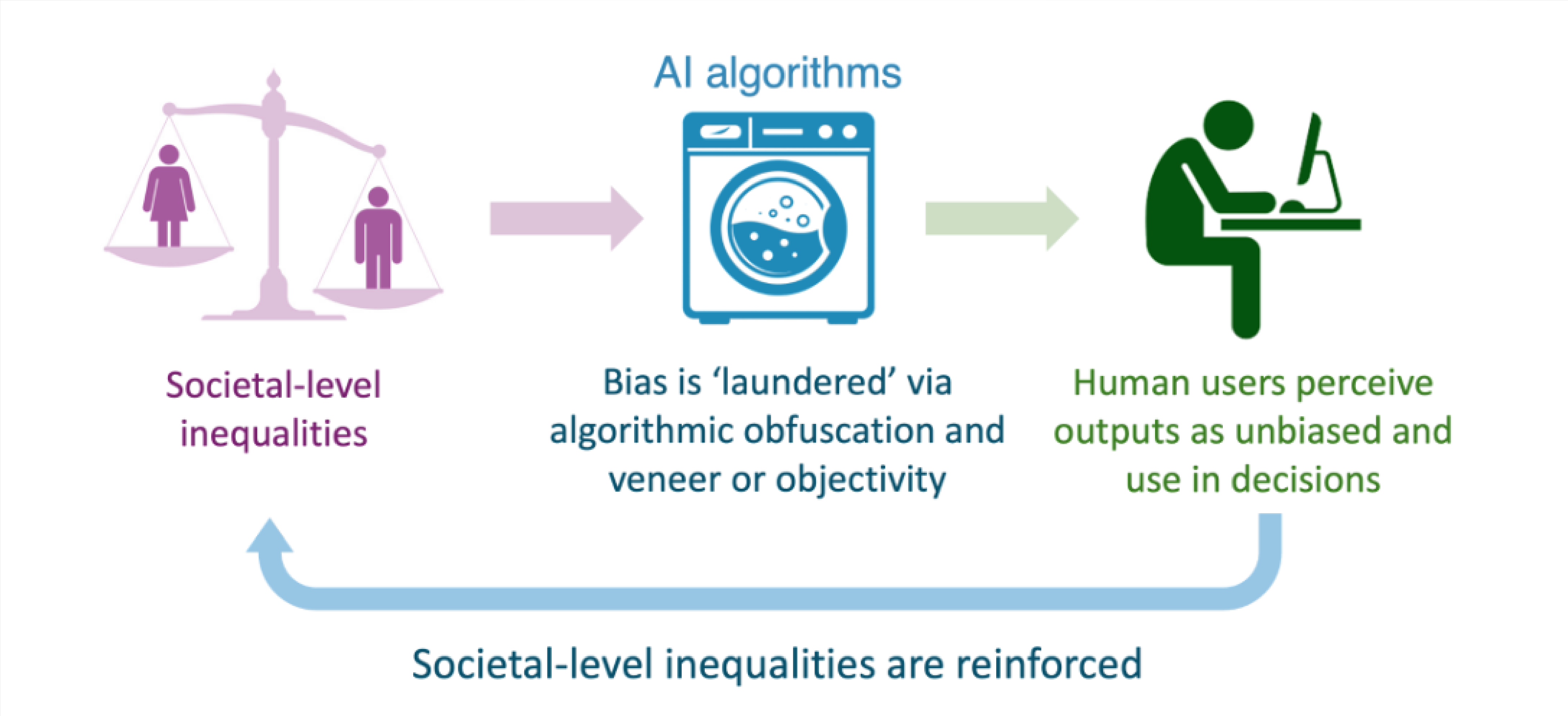
- By removing names?
  - No
- By providing more evidence about employee performance?
  - No
    - Also why not just write it yourself
- By asking the tool to avoid bias?
  - No
- By asking the tool to just stick to what we've written?
  - No

Can we mitigate bias in AI evals?

Recommendations

- Consider the possibility that AI is going to create worse, and more biased, evals than managers
- Double down on your people: train and support managers

Bias acceleration machine



Amodio, Charlesworth, & Brady

Can we mitigate bias in AI evals?

Recommendations for organizations

- Understand the limitations of AI in performance reviews
- Make sure HR leaders are keeping a close eye on the evaluations to ensure bias isn't impacting certain groups
- Insist on AI tools that allow HR leaders to draw aggregate insights and look for patterns

Can we mitigate bias in AI evals?

Recommendations for managers

- Pre-commit to what's important in the evaluation, and make sure you have evidence from the evaluation period to back it up
- Understand how bias can play out in performance evals so you are equipped to spot it
- If possible, run the prompt again using another name to compare the feedback

The harms of genAI

- Trustworthiness
- Inequality, marginalization, and violence
- Concentration of authority
- Labor
- Ecosystem and environment

Evaluating the Social Impact of Generative AI Systems, Solaiman et al, 2025
Construction and Consequences: The Human Impacts of Artificial Intelligence Data Centers, Bush, 2025

Cognitive implications of AI use

- LLMs offer immediate convenience but have potential cognitive costs
- In an experiment comparing LLM users to others:
  - LLM group performed worse than brain-only users at all levels (neural, linguistic, and scoring)

Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task, Kosmyna et al. 2025

How do we move forward from here?

- If your organization has leaned into AI performance evaluations and now wants to go back to humans:
  - Set clear expectations
  - Provide support for managers
  - Consider implementing training
  - Use free resources from biasinterrupters.org

Q&A