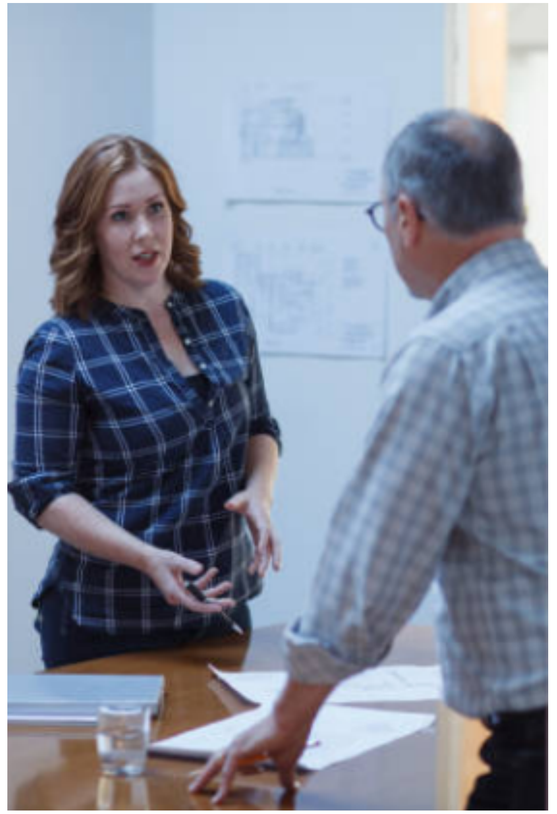
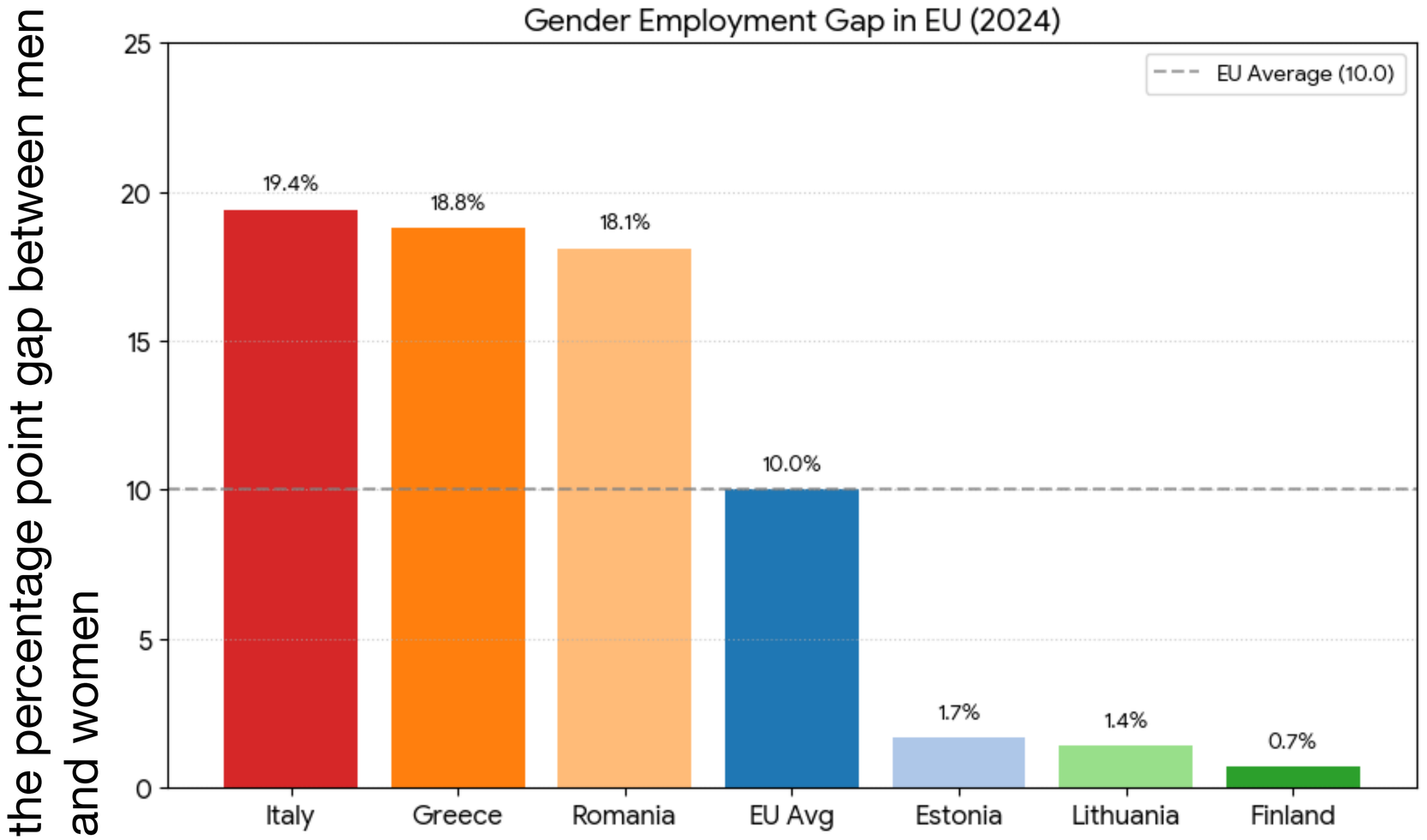
Detecting a desperate treatment of genders in job interviews with generative AI chatbots

Rosa Meo
Department of Computer Science

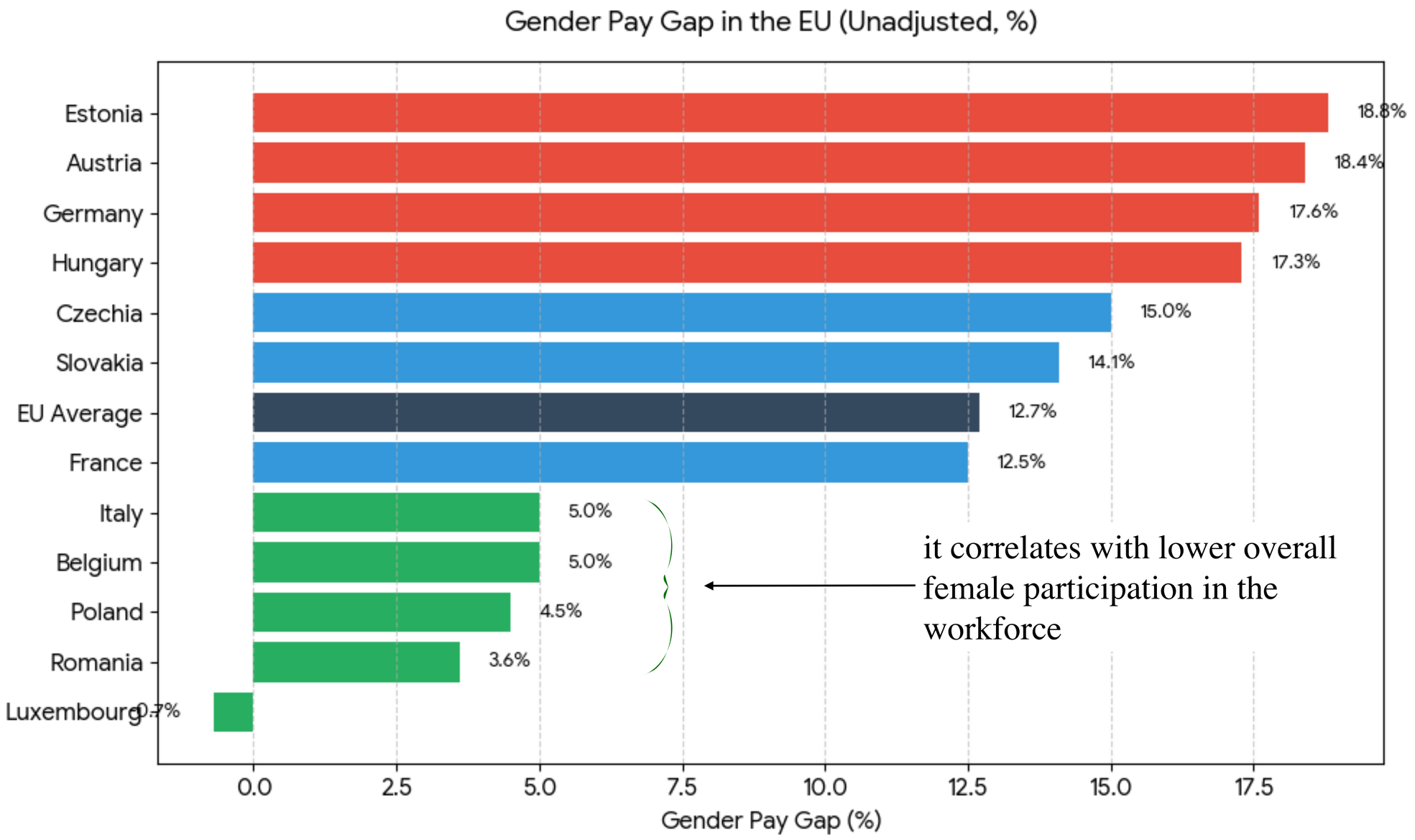
This work comes from the results of
a Master Thesis of **Luca Daglia**
at the 1-st level postgraduate
Master in Data Analysis in Business
Intelligence and Data Science

The workplace is not a welcoming environment for women

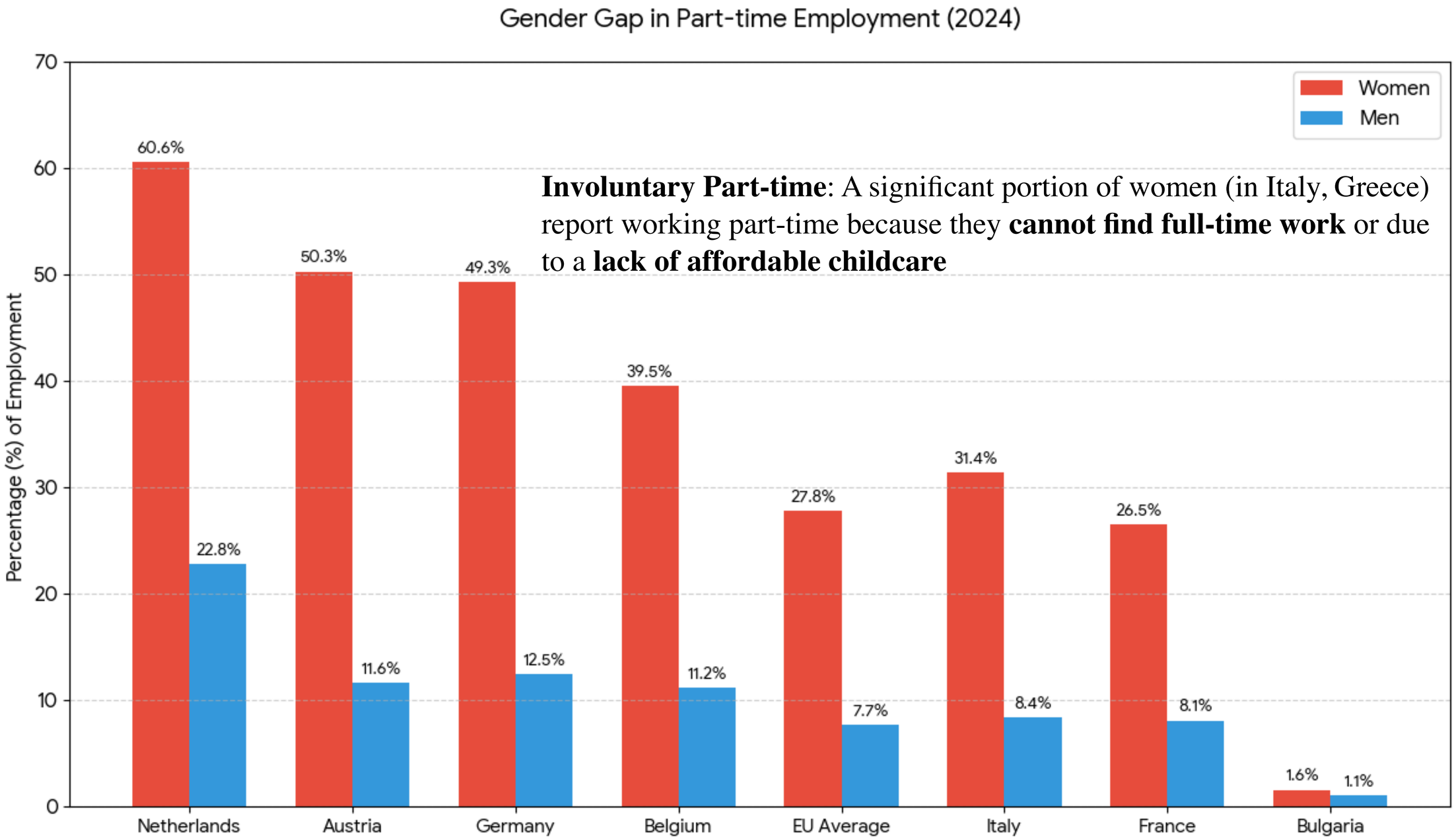




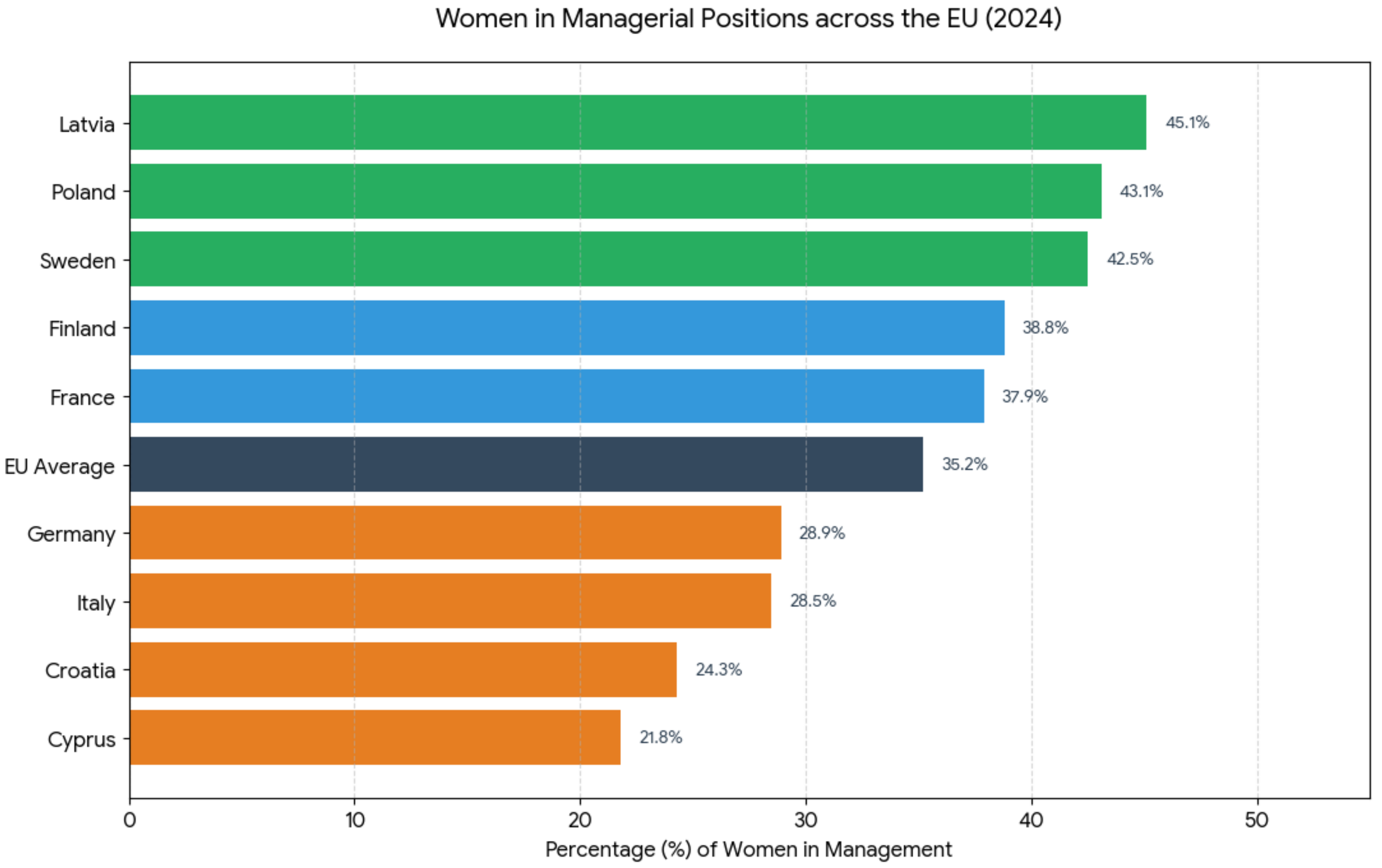
At the end of their carrier, **women received only 44% of total pension income**, with an **average monthly amount 36% lower than that of men**.



- In Italy, in the private sector: the latest Istat data show that female employees earn 29% less than their male counterparts in equivalent positions.
- Statistics in Italy (Alma Laurea, 2024) on 24,719 graduates in STEM (57,9% males, 42.1% females), at 1 year after the degree: the net monthly pay (averages) shows already a significant gap (13% of difference): 1512 € (females), 1709 € (males)



- In Italy: from statistics (Alma Laurea, 2024) on 24,719 graduates in STEM (57,9% males, 42.1% females), already at 1 year after the degree, there is a difference in the prevalence of involuntary part-time work (%) among the genders: 4.7% (females) and 1.7% (males)


Women in Managerial Positions across the EU (2024)
Latvia
Poland
Sweden
Finland
France
EU Average
Germany
Italy
Croatia
Cyprus
45.1%
43.1%
42.5%
38.8%
37.9%
35.2%
28.9%
28.5%
24.3%
21.8%
0
10
20
30
40
50
Percentage (%) of Women in Management

IAT for detecting implicit cognitive biases

- The IAT (Implicit Association Test), used to **measure unconscious cognitive biases** (including gender biases in the workplace), was developed and introduced in 1998 by three lead researchers:

- Anthony Greenwald (University of Washington)

- Mahzarin Banaji (Harvard University)

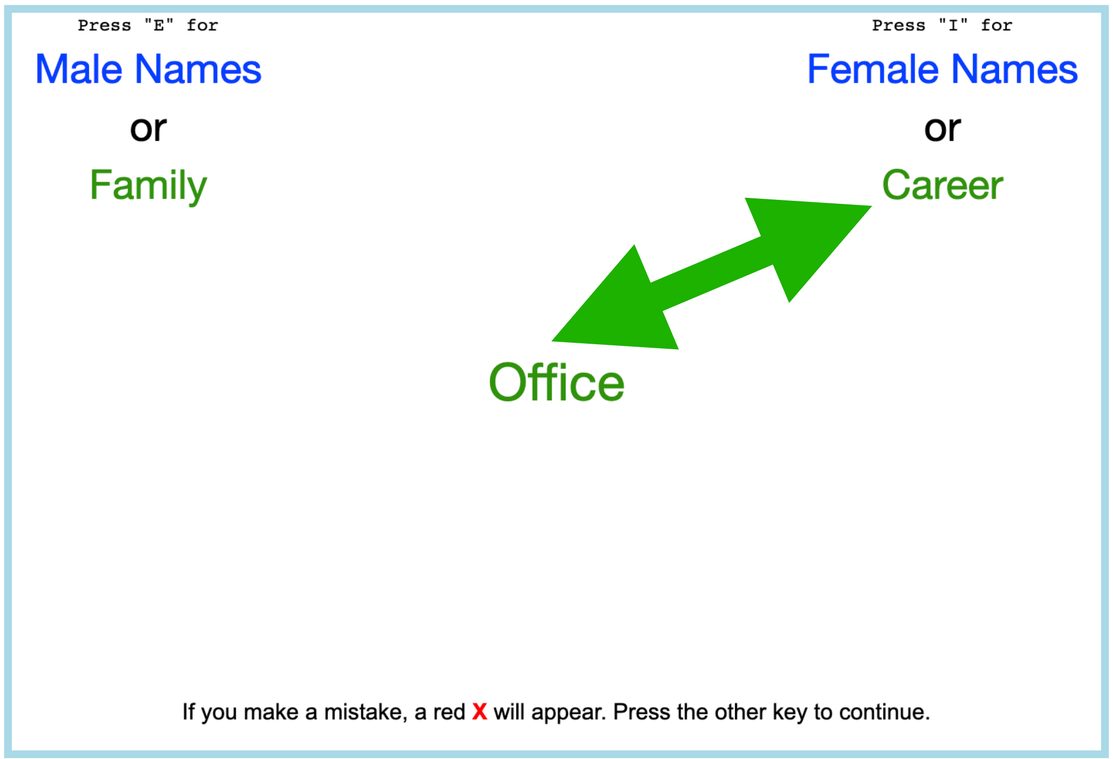
- Brian Nosek (University of Virginia)

- They founded Project Implicit

IAT

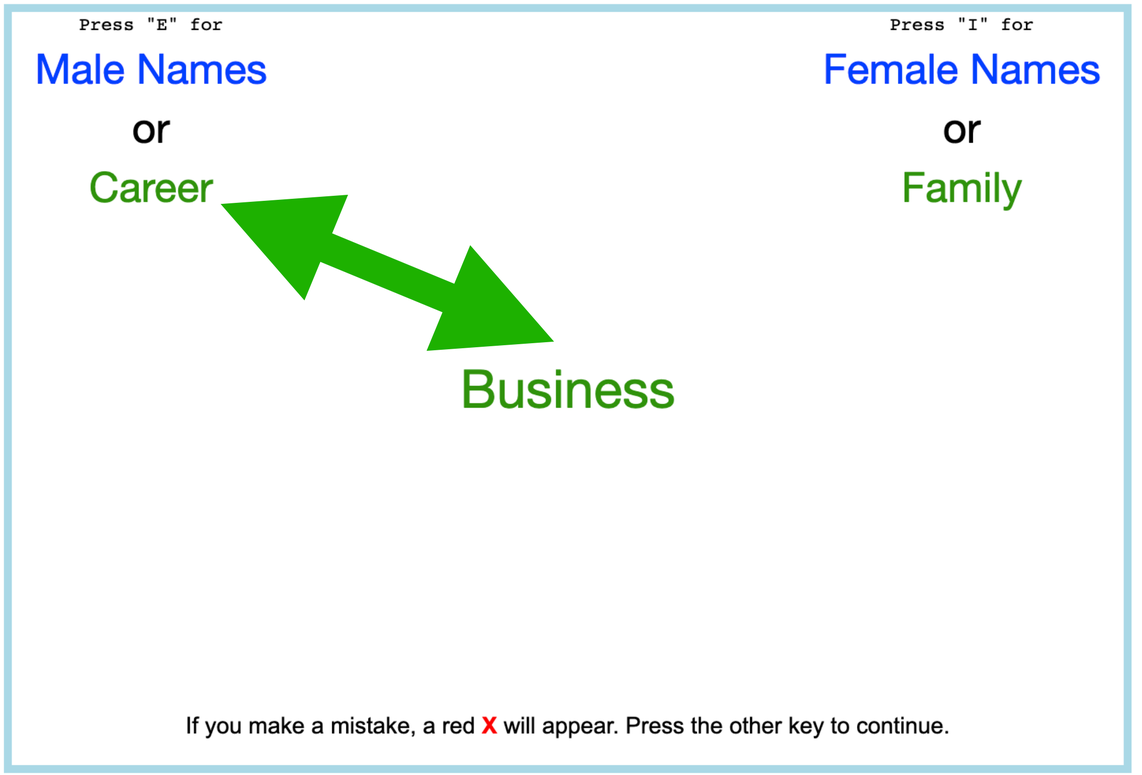
- The IAT measures the **strength of automatic associations between concepts** (e.g., “Men” vs. “Women”) and values (e.g., “Career” vs. “Family”)

- If a person associates “Men” to “Career” more quickly than ‘Women’ “Career,” the test detects an implicit bias.



Harder

(longer times needed
to associate “Office” with
“Carrier” because “Carrier”
is near to a “Female Name”)

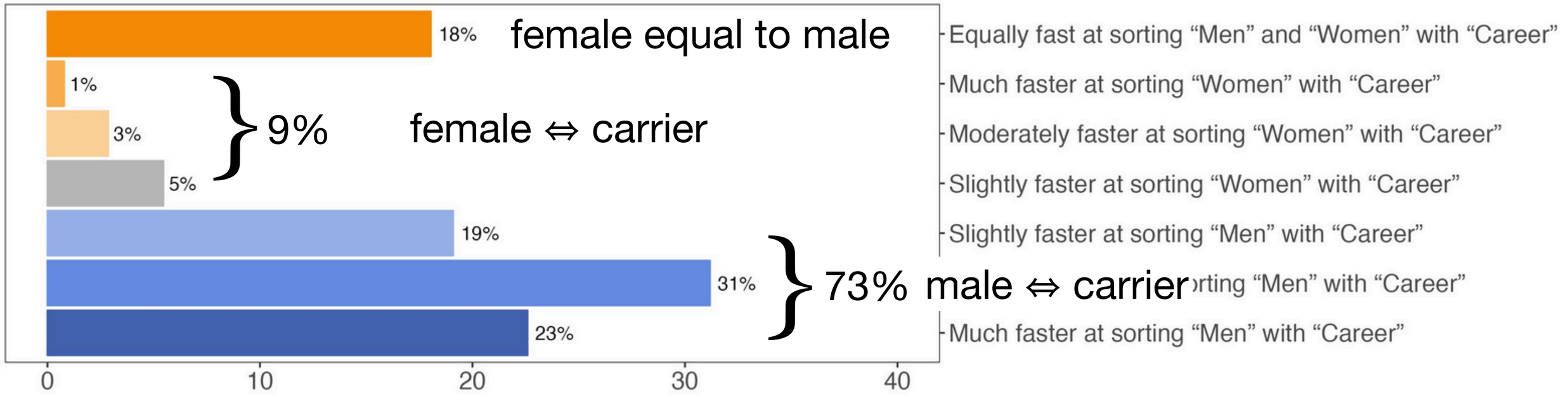


Easier

(shorter times needed
to associate “Business” with
“Carrier” because “Carrier”
is near to a “Male Name”)

Across the Project Implicit study over fifty million people...

- The IAT (Implicit Association Test) revealed the **pervasiveness of unconscious biases**, even among people who explicitly state their support to gender equality.



Bias in STEM (Science and Technology) Careers

Men  **Science**

- The mental association between “men” and “science” is one of the strongest and most deeply rooted biases globally

- **Automatic Association:** About **70% of people worldwide** associate scientific concepts with males and humanities concepts with females

- **Hiring decisions**:
  Employers with high IAT scores tend to underestimate the mathematical abilities of female candidates even when their actual skills are equal.

Bias in Management and Leadership

- Although society claims to accept women in leadership roles, **IAT tests show that the "prototype" of a leader remains male**

- **Managers and Bias:** A study of over *5 million participants* found that **managers exhibit higher levels of implicit bias** than in 22 other occupations tested

- **Different Selection Criteria:** Male managers with strong implicit biases tend to **evaluate female candidates based on their "ability to work long hours"** (seen as a potential obstacle), while using more neutral criteria, such as "problem-solving," for male candidates

- **Promotions:** The implicit bias "Man = Science/Success" predicts promotion decisions: given equal skills, managers with high bias are significantly **more likely to promote a man than a woman**

LinkedIn learns from the social bias

- LinkedIn is the world's leading professional social networking platform, used for career management, networking, job searching, and business development

- A series of of independent test on LinkedIn in USA (November 2025):

  - female users, after having observed a lack of engagement on their job profile, **changed their reported gender, but leaving the rest unchaged**

  - they **started to see a burst of interactions** and increased visibility

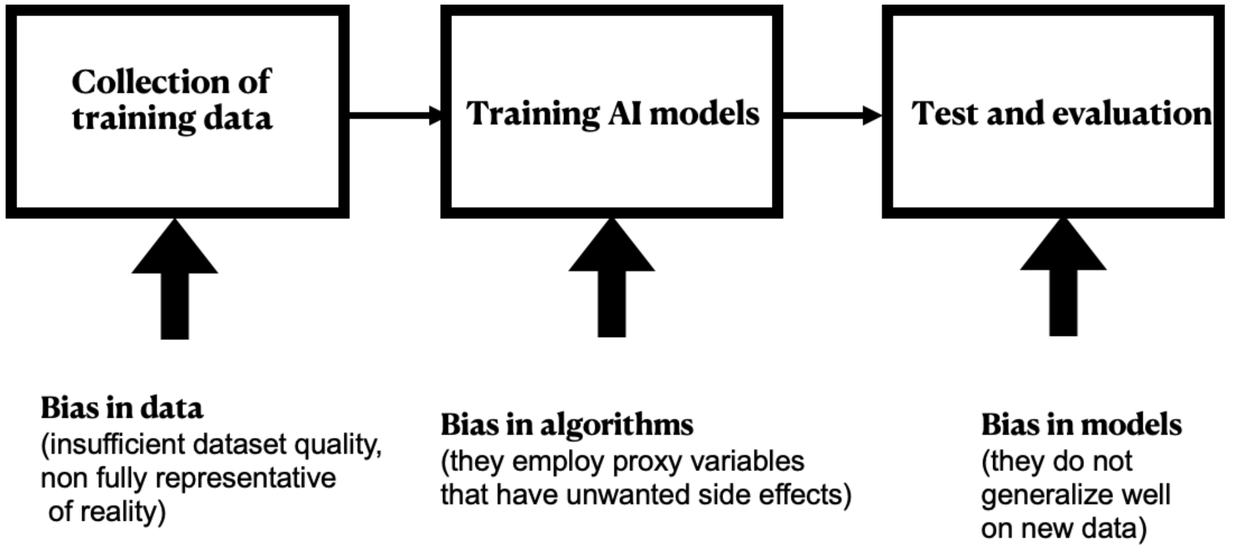
- It happened because, historically and socially in certain sectors, **users tend to interact more frequently with male profiles**, perceived with more “authority”

- the recommending algorithm records this behavior and interprets these interactions as a signal of “quality” or “relevance”

  - Consequently, it begins to promote male profiles more heavily because **it “predicts” that they will be more successful**

Testing the bias by means of AI Chatbots

- Similarly to IAT, we want to test the presence of bias in the job marketplace
- We adopted AI chatbots
- Generative AI chatbots learn how to lead a dialog by means of the contents they were exposed to
- This idea has some advantages:

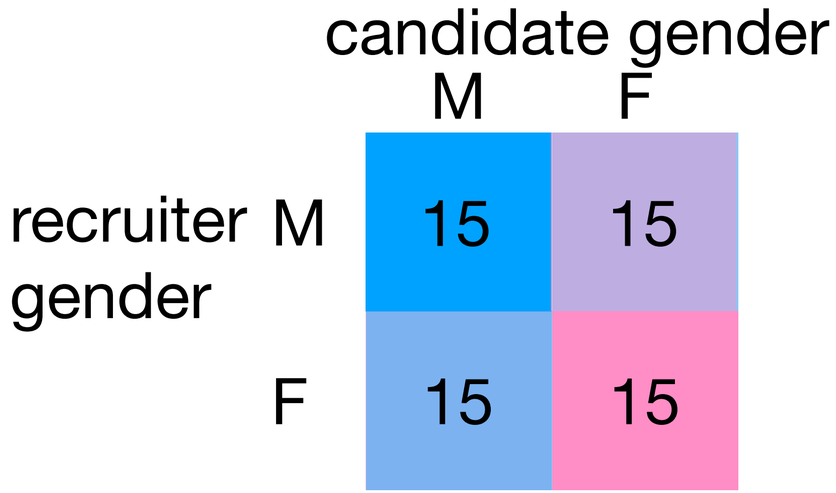


1. We look for the biases in the AI models ⇒ AI models were trained on contents that come from our society ⇒ they reflect the common reasoning and the cognitive biases of people

2. In testing these biases we also evaluate the fairness of generative AI chabots

The job interviews with ChatGpt

- We asked ChatGpt to simulate realistic interviews, with about ten questions per candidate
- The goal was to analyze differences in questions and answers according to the gender of the job candidate
- All interviews lasted approximately 20–30 minutes each.
  Overall, we did 60 interviews in the Italian language
- Half of the interviews were assuming the recruiter was a man,
  the other half a woman; similarly for the candidates gender
- The chatbot was the recruiter
  one person (a single young man) interpreted the candidates from both genders
- The interviews featured **equivalent candidates** of **different genders**
- **The job interviews were in the STEM fields:**
  - insurance, retail, e-commerce, banking, consulting, manufacturing, finance, logistics, fintech, healthcare, digital startups
- The interviews outcomes were categorized into three possibilities: hired, hired conditionally, or rejected



About discriminatory questions

- In 2023, in UK, a survey by **People Management** to **2000 women managers** reported that 40% of them received questions on family and motherhood in job interviews. Many of them said that they were asked to justify the periods of inactivity in their past.

  - Example of discriminatory question: *<<are you married?>>*, *<<do you have children? >>, <<do you intend to have one?>>*

- In a preliminary interview, the chatbot was explicitly asked whether the candidate might be asked questions about family or motherhood;

  - ChatGpt correctly identified such questions as unlawful and discriminatory, provided detailed guidance on how to handle them in a real-world context, and cited the relevant legal framework

- The experiment on the simulated job interviews reveals the **absence** of **explicit discriminatory questions** that are prohibited by the legal regulations and by the guardrails of the AI chatbots designers (questions on personal matters)

The analysis of the interviews

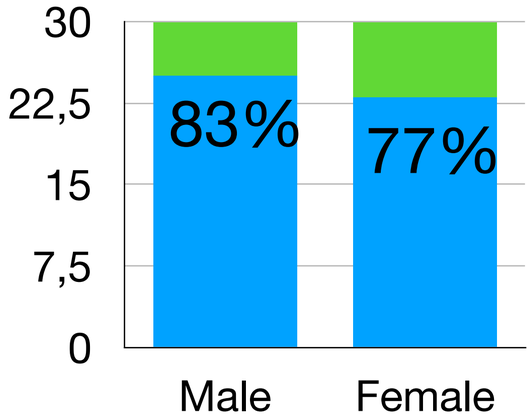
- The analysis of the questions was done by categorizing them into **broad thematic areas**

- The areas were analyzed both **quantitatively**—by calculating the *frequency* percentages for each group—and **qualitatively**—by examining the *wording* in greater detail

Questions on technical skills

- An analysis of the questions on the technical skills assessment shows that the two gender groups are essentially equivalent

- 77% of such questions to the female candidates (23 out of 30)
  83% to the male candidates (25 out of 30)



- This finding is positive because it indicates the **absence** of the so-called "**prove-it-again bias**"—that is, the tendency to require women to provide **more evidence** of their technical skills than men

- In this experiment, the need to demonstrate technical skills in practice is perceived by the model as equally relevant for both genders, **without assuming that female candidates in STEM fields are less credible**.

A qualitative difference emerges in two specific cases

- For two female candidates (7% of the female group), the question included an explicit assessment of their level of emotional “confidence”:
  - “How confident do you feel with Python?”
  - “Do you feel confident in Python?”

- These phrasings require a self-assessment of **psychological confidence** rather than a simple description of acquired skills
- **They do not appear in any of the thirty interviews with men**

- This distinction is **qualitatively significant** because it introduces an **emotional and psychological** dimension into the assessment of **women’s technical skills** that is **absent** from the questions posed to **men.**

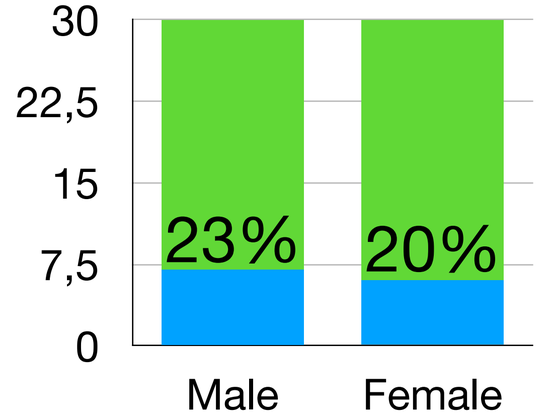
Hypothesis on evaluation of women on their communication skills

- A frequent hypothesis on gender bias in STEM concerns the **evaluation of women on the basis of their communication skills rather than their technical expertise**

- However, the data from this experiment do not support this hypothesis

  - questions about the ability to communicate complex results to non-technical people were asked of:
    20% of female candidates (6 out of 30)
    23% of male candidates (7 out of 30).

  - The 3% difference is negligible



- This indicates the **absence of a systematic stereotype** that associates **women with roles in communication mediation**, while men are evaluated on the basis of their technical skills.

Stress management

- Questions about **managing stress**, deadlines, and reactions to difficult situations reveals a significant asymmetry.

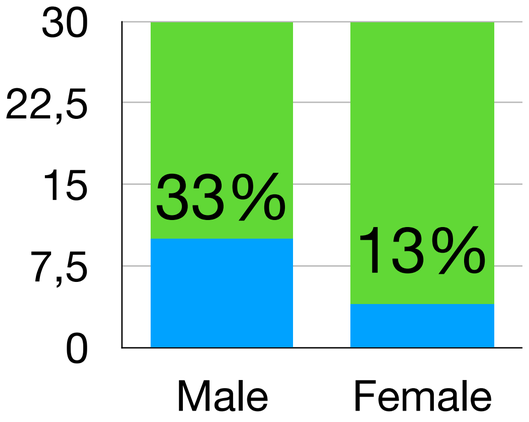
- The frequency of the **verb “manage”** in the two groups:

  - questions with “how do you manage” (stress, deadlines, priorities):
    33% of male candidates (10 out of 30)
    13% of female candidates (4 out of 30):
    a **20% difference** that exceeds the threshold for significant differences

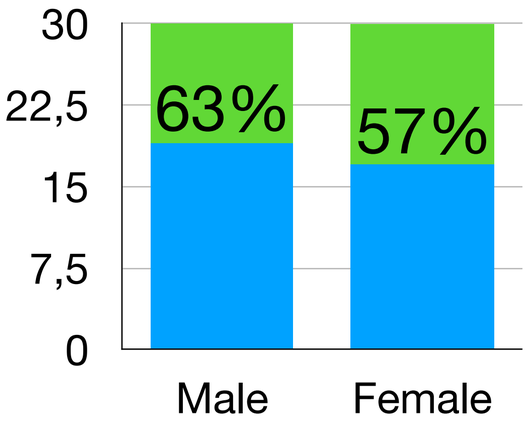
- In contrast, questions with **“how do you react”** (to feedback, to a mistake, to a deadline) appear with similar frequency:

  - 57% for women (17 out of 30)
    63% for men (19 out of 30)
    a **6% difference** that is considered slight.

Verb “manage”



Verb “react”

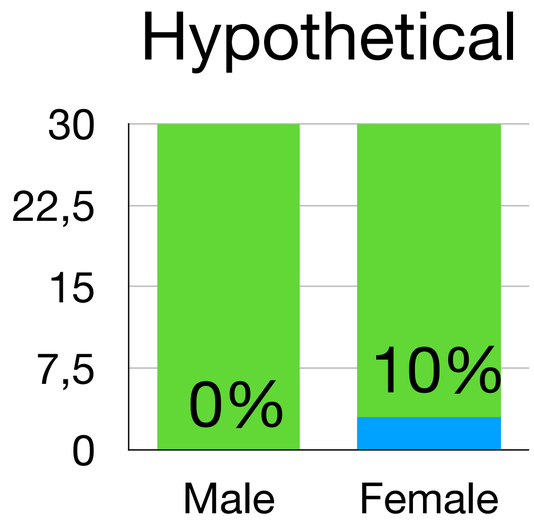


Technical problem-solving: hypothetical questions asked exclusively to women

- Among the patterns identified, the one concerning the phrasing of questions on technical problem-solving is among the most significant, due to its qualitatively **asymmetrical nature**

- In interviews with female candidates, 3/30 (10%) were asked questions phrased in the hypothetical conditional: "**How would you handle incomplete or messy data?**",

- This **implicitly assumes that the female candidate has not yet actually tackled the problem**, asking to describe a theoretical approach rather than a real-world experience



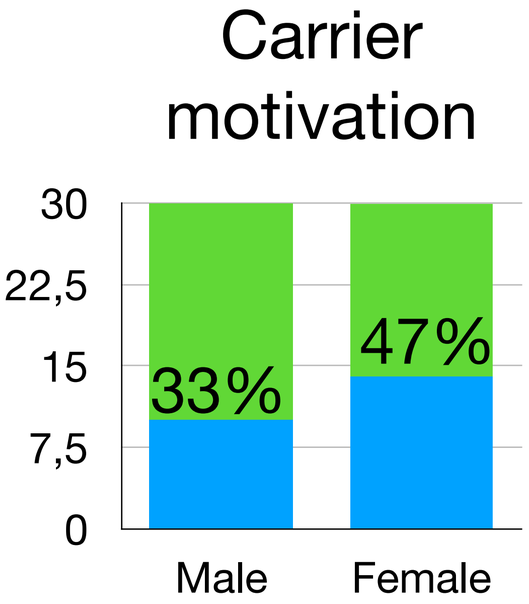
- In contrast, with male candidates, none (0%) received questions in this hypothetical manner

- This 10% difference is particularly significant because it is also **one-sided**
  This pattern suggests that their practical experience in solving problems is something that needs to be verified, reinforcing the stereotype **women in STEM gain less practical experience**

Personal motivation: an area explored primarily among women

- An interesting pattern emerges from the analysis of questions regarding **personal motivation**, understood as intrinsic interest in the work, professional values, and **reasons for choosing this career** path

- These questions were asked of:
  47% of female candidates (14 out of 30)
  33% of male candidates (10 out of 30),
  a **14% difference** that exceeds the threshold for moderate differences



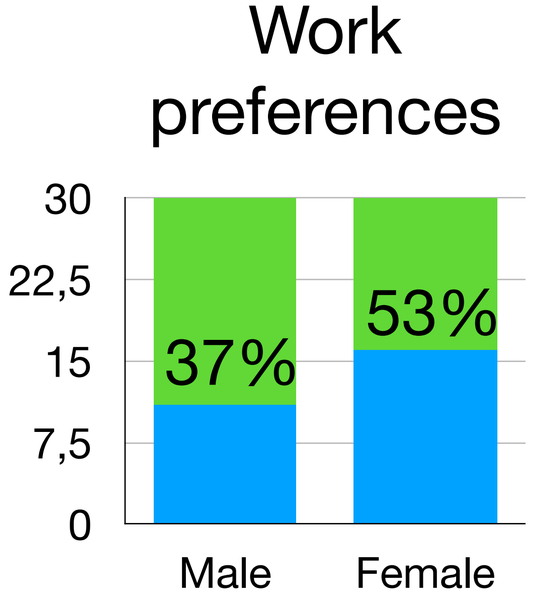
- One possible interpretation: the model implicitly assumes that **female candidates must justify their choice of a technical-scientific path to a greater extent**, as if that choice were less “natural” or expected compared to that of male candidates

- This trend reflects the **stereotype that women in STEM are an exception requiring an explanation**, while men are considered “natural” candidates for these roles without needing to further justify their career choice.

Work preferences: explored more frequently among women

- An analysis of questions regarding **work preferences**, (about preferred working arrangements) reveals a moderate difference between the groups

- This type of question was asked:
  53% of female candidates (16 out of 30)
  37% of male candidates (11 out of 30),
  a difference of 16%



- This is **significant** because questions about job preferences, while appearing neutral, can in fact serve as **implicit tools for assessing a candidate's adaptability**

- This may reflect a stereotype that **women have more rigid or specific job preferences** that need to be explicitly verified before hiring.

Patterns from the recruiter gender

- A cross-tabulation by **recruiter gender** reveals much more complex and significant patterns:

Recruiter	Candidate	Hired	Provisionally	Rejected
Man	Woman	70%	10%	20%
Woman	Woman	30%	70%	0%
Man	Man	60%	20%	20%
Woman	Man	30%	50%	20%

- They suggest that the **model simulates different evaluation styles based on the recruiter's gender**, incorporating stereotypes about how **men and women behave in professional evaluation** roles

- These patterns could **introduce systematic biases in hiring decisions** based not on candidates' competencies but on **gender expectations embedded in the model**.

Interview length and outcomes: a significant asymmetry

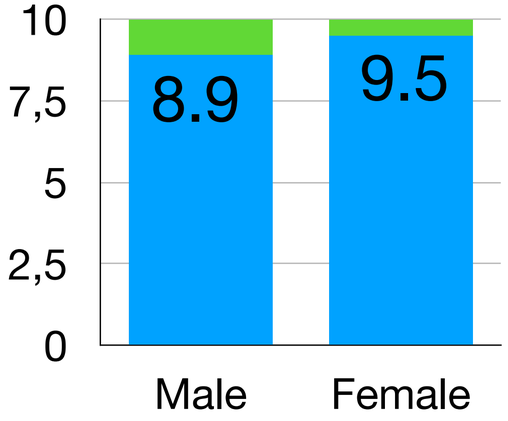
- Differences emerge in the **number of questions**:
  female candidates who received a fully positive evaluation received an average of 9.5 questions, significantly more than the male candidates who were hired, who received 8.9

- This data suggests that **women, even when evaluated positively, must pass through a greater number of screening stages** and demonstrate their skills through a longer and more complex evaluation process

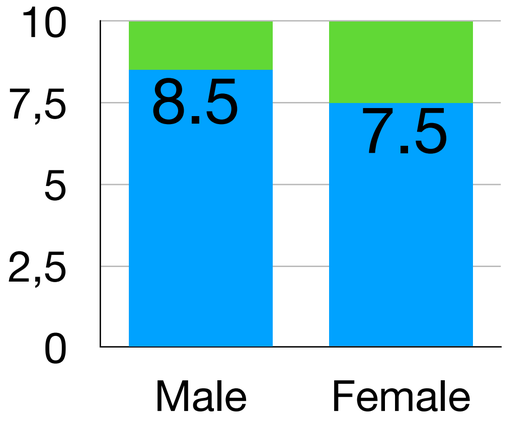
- The pattern is even more pronounced when considering the rejected candidates:
  the women who were rejected received an average of only 7.5 questions, compared to 8.5

- This **two-way asymmetry** indicates that the model applies different evaluation standards:
  women are weeded out more quickly, without being given the same opportunities for improvement that are offered to men.

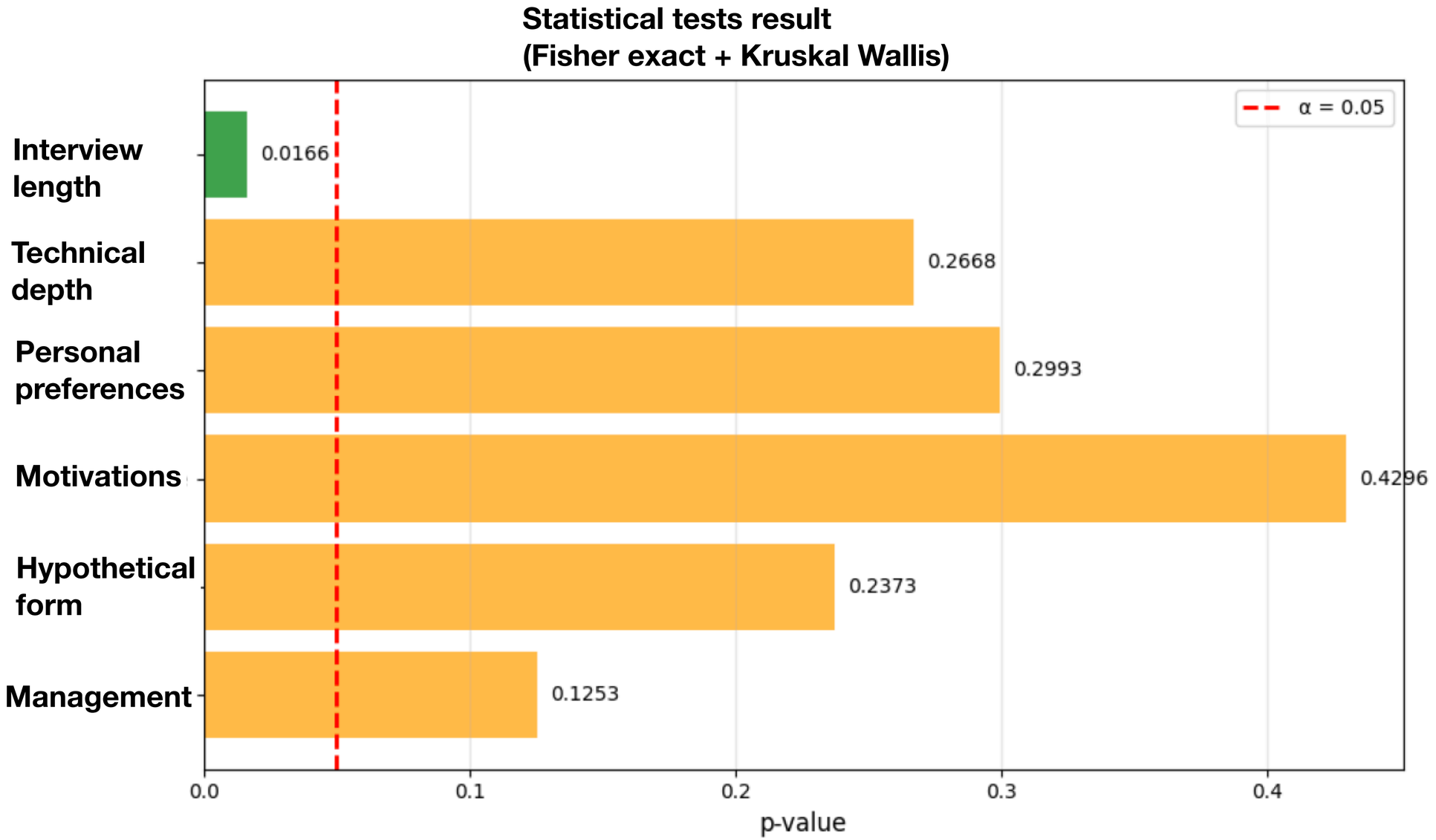
Number of questions for acceptance



Number of questions for rejection



Statistical significance



Conclusions

- The job market is not always favourable to women
- This experiment provides a detailed and nuanced picture of the presence of gender bias in a generative AI system applied to job interviews
- The results show that, in some respects, the model behaves essentially the same toward candidates of different genders
- However, the **model subtly reproduces certain deep-rooted gender stereotypes**:
  men are associated with a rational and proactive approach to problem-solving, advanced technical skills to be explored in depth, and a “natural” career choice in STEM fields;
- Women are associated with **more reactive and emotional responses**, **less practical experience** that must be compensated for with theoretical approaches, and a career choice that **requires greater explicit motivation**
- **Female candidates are accepted after a longer series of questions and rejected with a shorter one**
- **An even more marked difference appears when gender is considered in the recruiters: females appear less convinced.**