DEPARTMENT
OF POLITICAL AND
SOCIAL SCIENCES

Artificial intelligence, gender and work

Elena Pisanelli

Thesis submitted for assessment with a view to
obtaining the degree of Doctor of Political and Social Sciences
of the European University Institute

Florence, 09 November 2023

European University Institute
Department of Political and Social Sciences

Artificial intelligence, gender and work

Elena Pisanelli

Thesis submitted for assessment with a view to
obtaining the degree of Doctor of Political and Social Sciences
of the European University Institute

Examining Board

Prof. Klarita Gërxhani, Vrije Universiteit Amsterdam (formerly EUI, Supervisor)
Prof. Arnout van de Rijt, EUI
Prof. Chiara Monfardini, University of Bologna
Prof. Paola Profeta, Bocconi University

Researcher declaration to accompany the submission of written work
Department of Political and Social Sciences - Doctoral Programme

I Elena Pisanelli certify that I am the author of the work Artificial intelligence, gender and work I have presented for examination for the Ph.D. at the European University Institute. I also certify that this is solely my own original work, other than where I have clearly indicated, in this declaration and in the thesis, that it is the work of others.

I warrant that I have obtained all the permissions required for using any material from other copyrighted publications.

I certify that this work complies with the Code of Ethics in Academic Research issued by the European University Institute (IUE 332/2/10 (CA 297).

The copyright of this work rests with its author. Quotation from it is permitted, provided that full acknowledgement is made. This work may not be reproduced without my prior written consent. This authorisation does not, to the best of my knowledge, infringe the rights of any third party.

I declare that this work consists of 34,226 words.

I confirm that chapter 4 was jointly co-authored with Prof. Arthur Schram and I contributed 50% of the work.

Signature and date:

October, 11, 2023

Elena Pisanelli

Artificial intelligence, gender and work

Elena Pisanelli

Everyone wants to live on top of the mountain, but all the happiness and growth occurs while you are climbing it.

Andy Rooney

Abstract

This thesis explores the use of artificial intelligence (AI) technologies in hiring and their impact on gender inequality in the labor market. While AI has been adopted by firms with the expectation of unbiased decision-making processes, existing research shows that AI often tends to exhibit bias learned from humans, thereby reinforcing gender disparities. The thesis investigates why this is the case, focusing on two widely adopted AI tools in hiring: predictive algorithms and assessment software. Previous studies have primarily focused on predictive algorithms, demonstrating that they can perpetuate human biases due to their reliance on firms' historical employment choices to make the hiring decision. In contrast, assessment software, which evaluates candidates through resume screening or cognitive tests, has received less attention in the literature. This thesis sheds new light on how assessment software and predictive algorithms differently affect gender inequality in hiring.

Chapter 1 introduces the key aspects of the thesis, including its motivation, theoretical framework, and empirical strategy. It discusses the use of AI in hiring and highlights the fresh research perspectives it brings to the study of gender discrimination in the labor market. The chapter aims to (i) provide an overview of the thesis's contribution to the existing literature on AI and gender discrimination and (ii) explain the primary theoretical and empirical approach employed in the thesis.

Chapter 2 presents empirical evidence based on data from Global Fortune 500 firms, employing a difference-in-differences approach. By examining the combined impact of assessment software and predictive algorithms, the chapter shows that the use of AI in hiring increases the proportion of female managers hired by firms and is correlated with a reduction in firms facing gender discrimination lawsuits related to hiring.

Chapter 3 delves deeper into the study of AI and explores the heterogeneous effects of assessment software and predictive algorithms on gender inequality when they automate the hiring process. An intervention study conducted in a private company shows that, when granted full autonomy in hiring, assessment software significantly increases the representation of female applicants shortlisted for job interviews. Conversely, predictive

algorithms do not differ significantly from human recruiters in promoting gender diversity in the hiring process. Both AI tools ensure, unlike human recruiters, that the selected applicants are highly qualified.

Chapter 4 completes the picture by examining the use of assessment software and predictive algorithms as complements to human recruiters in the hiring process. By modeling employers' hiring choices and conducting an online experiment, the study demonstrates that both assessment software and predictive algorithms enable recruiters to escape information cascades. Additionally, both AI tools improve the overall productivity of selected applicants. Assessment software can also alter employers' prior beliefs about job candidates and enhance the diversity of hires, particularly when significant productivity differences exist among the job applicants under consideration.

The thesis concludes by emphasizing the importance of understanding the implications of assessment software's autonomy in hiring decisions. Although assessment software can reduce gender inequality in hiring when it automates hiring decisions and supports human recruiters, it may not address existing gender biases effectively when used in tandem with recruiters, similar to predictive algorithms. The thesis suggests that granting full autonomy to assessment software may be a more effective approach to reducing gender inequality in the labor market.

Acknowledgements

I would first of all like to thank my supervisor, Klarita Gërxhani. I am deeply grateful for having worked with her during these 4 years. I would not be here without her insight and motivation. During the Ph.D. journey, she was not only a supervisor but also a reference point, supporter, and mentor. Thank you, Klarita, for having bet on me. You have challenged me to be better, to be passionate about my research, and you have motivated me to push through.

I am also very grateful to Arthur Schram for helping me understand how to be analytical and for his guidance while working together.

This Ph.D. would not have been the same without the researchers and faculty of the European University Institute, who contributed to a fantastic atmosphere full of inspirational academic discussions and happy moments.

Indeed, a huge thanks goes to Team G. With your support and friendship, you have been fantastic companions on this journey.

I would also like to thank my mom, without whom this Ph.D. would never have been possible. She has always been my main supporter, and it is mainly thanks to her efforts and sacrifices that I can now be here accomplishing this goal and embracing my dream. I dedicate this thesis to her.

Last but not least, I would like to thank my partner, Giacomo, who has shared this journey with me. There are not enough words to express how grateful I am for having met you and for sharing my life with you. We not only completed our Ph.D. together, but we also made it to the wedding. Thank you for always being there for me.

Contents

List of Figures

List of Tables

1

Introduction

The pioneer computer scientist Alan Turing once expressed the wish that "machines would eventually compete and exceed men in all purely intellectual fields" (Turing, 1950, p.460).

This dream appears to be coming true. The use of artificial intelligence (AI) to solve complex problems has allowed us to exceed the computational limits of the human brain. In principle, AI should allow us to overcome another crucial limitation of human cognition: prejudice, defined as "an unfavorable opinion formed about a group without a basis in evidence" (Thagard, 2019, p.108).

But despite popular claims that AI overcomes human cognitive biases (Black and van Esch, 2020; Langenkamp et al., 2020; Bogen and Rieke, 2018), most scholars now argue that AI actually learns to be biased from humans (Gonzalez et al., 2022; Gebru, 2020; Köchling and Wehner, 2020; Black and van Esch, 2020; Silberg and Manyika, 2019; O'neil, 2017).

This raises the question: how can a technology that bases its reasoning on objectivity and evidence learn prejudices from humans? Does it depend on how different AI technologies acquire information, process it, and make decisions? How can these biases be overcome? And how much autonomy should be given to machines to make decisions that affect human lives? Addressing these questions will enable us to position AI within the realm of traditional socio-economic problems, such as gender discrimination. Furthermore, due to the significant interaction between AI and human cognition, studying whether and how AI might make biased choices can provide fresh insights into the role that human cognition plays in determining gender inequality. Because AI can expand the bounds of human rationality (Csaszar and Steinberger, 2022), studying how it differs from and interacts with human decision-making can shed light on the nature and limits of human rationality.

In this thesis, I address the case of the labor market, a typical case study of decision-making under uncertainty, where employers make choices with imperfect information regarding workers. Simon et al. (1990) define human behavior under uncertainty as bounded by human computational capabilities. Such limited capabilities lead humans to routinely violate axioms of probability theory and rational choice theory (Kahneman et al., 1982) when making decisions. This is because they employ learned heuristics and rules of thumb that may be inaccurate or biased (Thaler and Sunstein, 2009; Kahneman et al., 1982). AI, thanks to its computational capabilities, overcomes such cognitive biases and, thus, should reduce biases in employers' choices (Williams, 2022; Raisch and Krakowski, 2021; Black and van Esch, 2020; Daugherty et al., 2019). But existing research shows that AI is biased as humans are (Gonzalez et al., 2022; Köchling and Wehner, 2020; Cowgill and Tucker, 2020). This puzzle gives rise to the central research question of the thesis: does AI increase or reduce gender inequality in the labor market, and why?

An initial examination of the literature may lead one to conclude that this puzzle has been conclusively addressed by prior research. This argument is seemingly supported by the cited studies (Gonzalez et al., 2022; Köchling and Wehner, 2020; Cowgill and Tucker, 2020), which posit that artificial intelligence (AI) may exhibit gender-based discrimination as a consequence of the inherent biases ingrained in the training data. Nonetheless, a more thorough and meticulous inquiry is imperative to assess the veracity of this apparent resolution. Notably, extant research has predominantly concentrated its focus on a specific subset of AI, namely predictive algorithms, while concurrently, a distinct type of AI, assessment software, has become prominent within the labor market. It is striking to observe that there exists a conspicuous void within the existing literature with respect to this particular type of AI. Consequently, the puzzle at hand, though previously addressed by existing research, endures as a complex and unresolved issue that warrants renewed scrutiny, particularly in light of the advent of this new software.

To answer this question, I distinguish between the two main types of AI used in the labor market setting: predictive algorithms and assessment software. Predictive algorithms predict who to hire based on the characteristics and performance of previously employed workers (Rhea et al., 2022). Assessment software analyzes vast amounts of data contained in the job application material with speed and efficiency *without relying on data about previous employees* (Li et al., 2021). The latter distinction is of crucial importance and so bears repeating: predictive algorithms use history of previous hiring choices and past employees' characteristics, while assessment software does not. I propose that the central puzzle of this thesis is explained by this distinction.

As emphasized in the above paragraph, humans are boundedly rational. Their cognition is constrained both at the stage of gathering data and at the stage when the data is analyzed and decisions made (Viale, 2021). When rationality is constrained and decision-making is risky and costly, as in the labor market setting, humans seek to cut the cost of decision-making along a number of dimensions (time, computational power, etc.). In this context, heuristics, *i.e.*, simple decision-making rules based on repeated experience, observation, intuition, or common wisdom, can be efficient tools for making choices. Such a reliance on heuristics can give rise to discrimination in the labor market (Viale, 2021). Employers infer the expected competence, value and quality of workers either by relying on known statistics about the group to which the worker belongs (e.g., women's average productivity is lower than men's in specific jobs) (Phelps, 1972), or on widely held cultural beliefs about the group's competence and value (Ridgeway, 2011). Such group level information is surrogate and can be biased, giving rise to gender discrimination and inequality in the labor market (for a review on the topic see Correll and Benard (2006)).

In this context, at the stage of gathering data on workers, both predictive algorithms and assessment software reduce the costs that employers experience in making decisions, yielding vast amounts of information about workers (Daugherty et al., 2019). The difference between predictive algorithms and assessment software arises at the stage when the data is analyzed, where the two types of AI relate to heuristics in two different ways.

First, predictive algorithms use the average information about the group to which a worker belongs to infer her individual competence, quality, and value (Rhea et al., 2022). For example, predictive algorithms use historical data on successful employees to predict which candidates are likely to thrive in a particular role, or they analyze resumes and applications to identify candidates whose qualifications and skills match successful job-holders within the company. It seems reasonable to infer that although predictive algorithms expand human knowledge, they do not provide employers with additional information on individual workers' quality. Rather, predictive algorithms use statistics to infer individual information as humans do with heuristic strategies. Second, assessment software infer individual workers' competence without relying on statistics but using its computational power to elaborate information on individual workers. For example, assessment software screens job candidates' resumes and evaluates them according to the requirements of the job position, or evaluates candidates through cognitive tests based on psychological metrics. The estimate of individual-level information on workers' quality that comes from assessment software is more accurate and reliable than employers' private information (Williams, 2022; Raisch and Krakowski, 2021; Black and van Esch, 2020; Daugherty et al., 2019).

Both predictive algorithms and assessment software have different functions in decision-making processes. They can either augment human knowledge and work together with humans (Huysman and De Wit, 2004; Alavi and Leidner, 2001), or they can automate the decision-making process by replacing humans (Agrawal et al., 2018; Agarwal and Dhar, 2014). When they automate the decision-making process, predictive algorithms may likely reproduce past human biases, with bias defined as the evaluation of workers based on prejudice, i.e., 'an unfavorable opinion formed about a group without a basis in evidence' (Thagard, 2019, p. 108). This is because they rely on statistical information to infer the quality of individual workers (Gonzalez et al., 2022; Köchling and Wehner, 2020; Cowgill and Tucker, 2020).We also know from existing research that predictive algorithms, even when designed to be blind to gender, still perpetuate gender-based discrimination. This occurs because these algorithms often identify patterns within data that enable them to infer the applicant's gender and link it to the gender of incumbent workers, thereby using it to predict a successful match for the job position. For instance, they link gender to specific educational backgrounds (STEM vs. literature degrees) or career trajectories (managerial vs. teaching roles). Such associations, rooted in historical biases, result in gender disparities in algorithmic predictions (Gebru, 2020; O'neil, 2017). Conversely, assessment software may have good chances to remove human bias by accurately estimating individual workers' quality (Williams, 2022; Raisch and Krakowski, 2021; Black and van Esch, 2020; Daugherty et al., 2019). Assessment software relies solely on evidence regarding workers' work experience and the information reported on their resumes or performed cognitive tests. Therefore, accounting for individual socio-economic status and the associated biases that notably determine school and career achievements and can influence individuals' job prospects, assessment software should not introduce any additional discrimination into the hiring process. This ensures a more equitable evaluation of workers. For the same reasons, when they augment human knowledge, it seems reasonable to think predictive algorithms may validate human biases. By contrast, assessment software has the possibility to reverse human biases, but only if such biases are not deeply rooted in human cognition. If such biases are deeply rooted, then employers may be overconfident in their prior beliefs and even if they receive unbiased and perfect information, they will deviate from it and remain overconfident in their initial position. Employers may, therefore, interpret new evidence in a way that favors their existing preferences and insufficiently change their beliefs in response to the new evidence (Viale, 2021). For instance, employers that receive evidence about a woman being successful in a leadership job can perceive the woman as running counter to gender stereotypes and punish her by not hiring her (Quadlin, 2018; Heilman and Okimoto, 2007; Rudman and Glick, 2001). This is because employers evaluate the new evidence in favor of their rooted cultural beliefs that posit women should not be agentic, rather communal. Hence, by studying how humans make decisions when assisted by predictive algorithms versus assessment software, we can learn something important about the roots of human bias. Moreover, we can learn how to overcome these biases and what the role of either type of AI therein is.

In the three empirical chapters of the thesis, I, therefore, focus on the stage when the data is analyzed and decisions are made during the hiring process. While the literature has focused on predictive algorithms, it has left open questions of how assessment software may affect hiring decisions. I, therefore, study how both types of AI heterogeneously affect gender inequality in the labor market, by considering the interplay between the mechanisms and functions of AI identified above. To do so, I answer the following research questions:

RQ1: What is the combined average effect of predictive algorithms and assessment software on gender inequality in the labor market?

RQ2a: How do assessment software and predictive algorithms heterogeneously affect the qualifications level and diversity of hires when they automate the hiring process (*i.e.*, replace human decision-making)?

RQ2b: Do assessment software, predictive algorithms and human recruiters differ in this respect?

RQ3a: How do assessment software and predictive algorithms heterogeneously affect gender inequality in the labor market when used together with human recruiters?

RQ3b: What is the impact of assessment software's and predictive algorithms' evidence on human recruiters' prior beliefs?

The rest of the introduction is structured as follows. Section 1.1 presents what we know from the literature on AI and why we should extend our knowledge about it. Section 1.2 summarizes the theoretical framework underlying the thesis. Section 1.3 provides an overview of the empirical chapters.

1.1 What we know and need to know about AI

AI is challenging traditional ways of managing businesses, including human resources (HR) practices, such as hiring (Bhatt, 2022). The existing literature shows that AI is most suitable to be used at the initial screening stages of hiring (for a review see Bhatt (2022)). The particular context in which AI technologies are employed, specifically in hiring during the initial screening stage, highlights the relevance of studying the impact of AI on gender inequality in the labor market. We know from existing studies on human-powered hiring that once women are included in the candidate pool, they are as likely as men to be hired after having passed the interview stage of hiring, but they are less likely of being shortlisted to be interviewed compared to men (Fernandez-Mateo and Fernandez, 2016). Such a gap at the beginning of the hiring process is partly explained by recruiters bending the pipeline against women (Fernandez-Mateo and Fernandez, 2016). Since AI is most suitable to be used at the initial screening stage of hiring, when human recruiters bend the pipeline against female applicants, AI has the potential to improve women's disadvantage in the labor market.

However, existing research prompts concerns about AI reinforcing discrimination and widening gender inequality in hiring (Gebru, 2020; Danieli et al., 2016). When AI automates hiring decisions, it reflects human biases that are embedded in the data that trains AI algorithms (Cowgill and Tucker, 2020; O'neil, 2017). The historical data that train AI algorithms can reflect existing social inequalities, which AI replicates, for example, by selecting candidates that reflect companies' historical hiring patters (Vasconcelos et al., 2018; O'neil, 2017). AI selects those candidates that are already favored in the labor market, reinforcing the existing biases and increasing inequality (Gebru, 2020). Moreover, when AI augments human knowledge in the hiring process, the interaction between humans and machines increases the risk of human biases being carried over to machines (Raisch and Krakowski, 2021). AI ends up confirming human biases and reinforcing gender discrimination and inequality in the labor market (Huang et al., 2012). In sum, the literature highlights that both when it is used for automating hiring decisions and for augmenting human knowledge in hiring, AI is likely to increase gender inequality in the labor market.

On top of assuming different functions in the hiring process, *i.e.*, automate hiring decisions or augment human knowledge, AI can also assume different forms, *i.e.*, predictive algorithms or assessment software. From existing research, we know about predictive algorithms that ”a hiring model attempting to predict the characteristics determining a candidate's likelihood of success would invariably learn that the under-sampled majority are unlikely to succeed because the environment is hostile towards [...] women" (Gebru, 2020, p.8). Because predictive algorithms are built by finding statistical patterns in the data, they suffer from statistical bias, meaning that predictive algorithms can use arbitrary features, such as gender, to make hiring decisions (Langenkamp et al., 2020). How does assessment software relate to bias and gender inequality? This other half of the AI picture is still missing. However, the study of how assessment software affects gender inequality in the labor market and its relationship with bias is fundamental. In fact, we know that, contrary to predictive algorithms, assessment software does not rely on historical data about firms' hiring choices, the main documented source of bias in AI. Rather, assessment software evaluates job applicants based on performance-based questions and compares the answers with the job requirements (Chandler, 2017). It also evaluates candidates based on neuroscience games that test candidates' behavior and skills (Houser, 2019). Assessment software, thus, allows to standardize the initial evaluation of job candidates and can potentially remove human biases from the hiring process and reduce gender inequality in the labor market.

Section 1.2 presents the theoretical framework that this thesis advances to disentangle the role of predictive algorithms and assessment software in hiring, contributing to the literature by answering to the puzzle: does AI increase or reduce gender inequality in the labor market?

1.2 Theoretical framework of the thesis

1.2.1 Gender stereotypes and discrimination

Gender stereotypes affect how people perceive themselves and their expectations of how others behave (Ellemers et al., 2018). They are beliefs related to the gender of individuals that shape expectations regarding people's behavior, suggesting that women and men behave differently because of their gender (Hentschel et al., 2019). Gender stereotypes, thus, influence expectations about, for instance, the ideal occupation men and women should occupy according to their role in the society (Eagly, 1997). Such expectations can give rise to gender biases and gender inequality in the labor market. Because employers have imperfect information about the future performance of job candidates, they have an incentive to rely on easily observable ascriptive characteristics, such as gender. Therefore, employers are influenced by gender stereotypes when inferring job applicants' expected performance if hired (Correll and Benard, 2006). In fact, both economists and sociologists have advanced theories to explain why employers rely on gender stereotypes and discriminate based on gender in hiring and promotion decisions.

Statistical theory of discrimination highlights a key explanation for why employers discriminate based on gender: the excessive cost of gaining information about individual workers' productivity leads employers to rely on group (men or women) statistics to infer information on individual workers (Arrow, 1973; Phelps, 1972). Gender discrimination, thus, arises as rational solution to an information problem.

Sociologists have advanced another explanation for why employers end up discriminating based on gender. Such an explanation finds its roots in expectation states theory (Berger, 1977; Conner, 1974), which assumes that employers discriminate against women because of their biased expectations of women's competence and hence women's future performance (*i.e.*, productivity) if hired. Contrary to statistical theory of discrimination, such biased performance expectations do not arise from statistical evidence but from biased *status* beliefs, defined as "widely held cultural beliefs that link greater social significance and general competence (*i.e.*, status), as well as specific positive and negative skills, with one category of a social distinction (e.g., men) compared to another (e.g., women)" (Ridgeway, 2001, p.638). These beliefs may affect gender inequality in the labor market through both demand and supply side mechanisms. This thesis focuses on the demand side, where the existing sex-segregated structure of the contemporary labor market makes cultural beliefs about gender salient (Ridgeway, 2011). The gender stereotype pertaining to the sex that predominates in a job biases the traits of the ideal worker for that job (Reskin and Bielby, 2005). For example, being a nurse is typically associated to female traits (empathy, compassion, caregiving, ...) and, thus, considered a typically female job, leading recruiters to prefer women over men for such a job. Further, the prestige associated with the job itself can make cultural beliefs about gender salient (Ridgeway, 2019). In the Western world, jobs associated with authority and competence are, in fact, usually culturally associated with masculinity (Powell et al., 2002). Cultural beliefs about gender can, thus, set the stage for gender discrimination in hiring by biasing employers' perception of men and women's competence to perform a given job (Ridgeway, 2011; Heilman and Okimoto, 2007; Gorman, 2005). More problematically, when women apply for authority positions or self-present as agentic and assertive, they are punished for it by employers (Quadlin, 2018; Heilman and Okimoto, 2007; Rudman and Glick, 2001). This is because incongruity arises between women's behavior in presenting themselves as agentic and cultural beliefs that presume lower status position and communality for women compared to men (Prentice and Carranza, 2002; Rudman and Glick, 2001). Such an incongruity leads to a backlash from the demand side of hiring.

1.2.2 New perspectives from AI

The emergence of AI provides new perspectives for the study of gender stereotypes, employers' beliefs and their impact on gender inequality in the labor market.

In the context of statistical theory of discrimination, predictive algorithms use the historical average information about the group to which the job applicant belongs to infer how the job applicant will perform if hired. More precisely, predictive algorithms hitherto have been programmed to estimate the average group (e.g., men, women) probability of success for firms' past workers, to then predict the future performance of individual job applicants belonging to the same group (Langenkamp et al., 2020). Thus, it seems reasonable to infer that how predictive algorithms behave in hiring is not any different from how human recruiters make their hiring decisions: they both use surrogate and biased information about the group to infer the individual job applicant's productivity. Theoretically, we should expect no difference between humans and AI in affecting gender inequality in the labor market, both when predictive algorithms augment human knowledge and when they automate the hiring process.

In the context of expectation states theory, predictive algorithms are not subject to status beliefs ("widely held cultural beliefs that link greater social significance and general competence, as well as specific positive and negative skills, with one category of a social distinction (e.g., men) compared to another (e.g., women)" (Ridgeway, 2001, p.638)), because, after all, they are a software. However, relying on data about firms' past employees, predictive algorithms associate gender to the probability of success in the job (Gebru, 2020; O'neil, 2017). Thus, as existing research suggests, predictive algorithms carry over employers' biased hiring decisions and biased beliefs (Gonzalez et al., 2022; Black and van Esch, 2020; Gebru, 2020; Köchling and Wehner, 2020; Langenkamp et al., 2020; Silberg and Manyika, 2019; Bogen and Rieke, 2018; O'neil, 2017). This may happen both when predictive algorithms augment human knowledge and when they automate the hiring process.

Consider now assessment software. In the context of statistical theory of discrimination, assessment software gains vast amount of information about individual job applicants' skills and performance, and uses such an information to select the job applicant to hire (Daugherty et al., 2019). Contrary to predictive algorithms and human recruiters, assessment software does not need to rely on surrogate and biased group information to infer individual workers' performance. Rather, it relies on and provides employers with accurate information on individual job applicants' productivity. Thus, it seems reasonable to infer that assessment software would not lead to statistical discrimination based on gender both when it automates the hiring process and when it augments human knowledge.

In the context of expectation states theory, as with predictive algorithms, assessment software is not subject to status beliefs when evaluating job applicants because it is a software and not a human being. Because it does not rely on data about firms' past employees, it seems reasonable to infer that assessment software should not carry over human biases both when it automates the hiring process and when it augments human knowledge.

In sum, as existing research suggests, we expect predictive algorithms to reproduce the existing gender inequality in the labor market, both when they automate the hiring process and when they augment human knowledge (Gonzalez et al., 2022; Black and van

Esch, 2020; Gebru, 2020; Köchling and Wehner, 2020; Langenkamp et al., 2020; Silberg and Manyika, 2019; Bogen and Rieke, 2018; O'neil, 2017). Conversely, it seems reasonable to expect that assessment software reduces gender inequality in the labor market, both when it automates the hiring process and when it augments human knowledge. The latter, however, needs further elaboration.

Assessment software is likely to reverse human biases, but only if such biases are not deeply rooted in human cognition. The process of stereotyping is a strategy that humans adopt for perceiving and making sense of the world (Allport et al., 1954; Lippmann, 1965). Since humans "cannot handle the complexity of their environment, they reconstruct it on a simpler model" (Lippmann, 1965, p.16). Such a simpler model relies on placing experiences, objects and people into categories that define the object and allow to extract meaning from the environment (Allport et al., 1954). Stereotypes represent beliefs associated with categories of people that help humans rationalize their behavior toward others (Allport et al., 1954). Once a person is placed within a category (e.g., men or women), stereotyping immediately takes place (Fiske and Taylor, 1991).

When additional information is available, people make use of such information and attach importance to it in different ways. If the information confirms the expectation people have regarding the category (e.g., women are communal), people assimilate the information into their existing (prior) beliefs and reinforce their stereotypes (Operario and Fiske, 2001). On the other hand, when the information dis-confirms the expectation people have regarding the category (e.g., a woman who is assertive), people tend to perceive the information as not representative of the category and find ways to interpret it in favor of their existing (prior) beliefs (Operario and Fiske, 2001). It seems, thus, reasonable to infer that in this scenario, employers receive the information coming from assessment software and interpret this new evidence in a way that favors their existing preferences (Viale, 2021). For instance, employers that receive evidence about a woman being successful in a leadership job can perceive the woman as running counter to gender stereotypes and punish her by not hiring her (Quadlin, 2018; Heilman and Okimoto, 2007; Rudman and Glick, 2001). However, when prior beliefs are weak or people are motivated to collect new information, perceivers can assimilate the dis-confirming information and modify their existing (prior) beliefs regarding the category (Hilton and Von Hippel, 1996). In the latter scenario, it seems reasonable to infer that employers receive the information coming from assessment software and may change their existing preferences accordingly.

In the three empirical chapters of the thesis, I, thus, study how predictive algorithms and assessment software affect gender inequality in the labor market. I then distinguish between when assessment software automate the hiring process and when they augment human knowledge, exploring how employers elaborate information that confirms or disconfirms their existing (prior) beliefs about job candidates. Section 1.3 presents the overview of the empirical chapters.

1.3 Overview of the empirical chapters

1.3.1 Chapter 2: a new turning point for women

This chapter answers to RQ1: 'What is the combined average effect of predictive algorithms and assessment software on gender inequality in the labor market?'

Empirical studies of labor market discrimination have a long tradition of methods and their applications. Simple regression analysis or its more complicated forms (e.g., difference-in-difference approach, instrumental variable approach, ...), with control variables that include observed productivity characteristics and a binary variable for race or gender, has historically been used to provide empirical estimates of gender inequality in the labor market.

In this chapter, I establish through a staggered difference-in-differences with ex-ante matching the causal relationship between firms' use of AI in hiring and gender inequality in managerial jobs. This approach allows for a rigorous control for unobserved confounders, which is especially important when the treatment is not randomly assigned, as with firms' use of AI in hiring. To do so, I use data on the largest companies by revenues in Europe and the US. The focus on such companies allows me to exploit vast amount of data, that are usually not available for small and medium sized firms, especially related to the use of AI. Further, such companies are the first movers in the use of AI in hiring, allowing me to exploit time variation to evaluate the impact of AI on gender inequality in the labor market.

I find that firms' use of AI causes, on average, a relative increase by 40% in the hiring of female managers. Also, I find that firms' use of AI correlates with a reduction in firms facing gender discrimination lawsuits related to hiring.

This chapter answers to the central research question of this thesis by highlighting that the impact of AI on gender inequality in the labor market is a complex and multi-faceted issue. AI can have both positive and negative effects on gender inequality, depending on how it is used and on the data it is trained on. On the one hand, assessment software can help to ensure that all applicants are given a fair and equal opportunity to be considered for a job because it evaluates job applicants based on objective criteria, such as their skills and qualifications, rather than subjective factors, such as their gender. On the other hand, predictive algorithms can perpetuate gender inequality in the labor market because they reflect employers past hiring choices.

Indeed, the most important limitation of this chapter is the fact that I measure the use of AI in hiring by collecting self-reported micro-level data on firms. After addressing potential endogeneity issues through my empirical strategy, this data allows me to estimate the aggregate effect of assessment software and predictive algorithms on gender inequality in hiring outcomes, by measuring the extent to which these AI tools affect the hiring process. However, I do not have information on whether firms use those hiring tools for automating hiring decisions, or augmenting human knowledge in hiring, or both. Further, I cannot estimate and compare with my data the heterogeneous impact of assessment software and predictive algorithms on gender equality.

In chapter 3 and chapter 4, I, thus, address this limitation by assessing respectively what is the heterogeneous effect of assessment software and predictive algorithms on gender inequality in the labor market (i) when they automate the hiring process, and (ii) when they augment human knowledge in hiring.

1.3.2 Chapter 3: the hiring dilemma: efficiency, equality or both?

This chapter answers to RQ2a: 'How do assessment software and predictive algorithms heterogeneously affect the qualifications level and diversity of hires when they automate the hiring process (*i.e.*, replace human decision-making)?' and RQ2b: 'Do assessment software, predictive algorithms and human recruiters differ in this respect?'

Besides quasi-experimental methods, empirical studies of labor market discrimination have long relied on intervention studies within companies (see, for example, Castilla (2015)). This method is often used in the evaluation of organizations' meritocratic practices, such as in evaluating merit-based reward systems (c.f. Castilla (2015)). Intervention studies allow for the direct manipulation of factors that may influence gender inequality in the labor market, as firms' use of assessment software or predictive algorithms in hiring. Therefore, using an intervention study allows to test the effectiveness of automating the hiring process through assessment software and predictive algorithms in affecting gender inequality. Additionally, by conducting the study within a firm, I can obtain a more realistic and applied understanding of how assessment software and predictive algorithms operate in the workplace.

In this chapter, I perform an intervention study within a private company. Through constructing a dataset of job applicants' resumes and their characteristics, including demographic information and gender, I compare the decision of whom to shortlist for the interview made by human recruiters alone, by assessment software alone, and by predictive algorithms alone.

This chapter views hiring as a balance between efficiency and equality: to find the best worker to hire, firms should balance the cost of screening and selecting job applicants' resumes with striving for gender equality in hiring outcomes. Indeed, modern hiring algorithms are designed for reducing costs. However, existing research shows hiring technologies based on predictive algorithms may well reproduce existing biases against women. I study how assessment software and predictive algorithms affect gender inequality in hiring when they automate the hiring process.

I show that assessment software increases the probability of selecting female applicants, compared to human recruiters and predictive algorithms. Further, I show that both assessment software and predictive algorithms increase to 1 the predicted probability to select highly qualified job applicants, compared to human recruiters.

My results answer to the central research question of the thesis by showing that when it automates the hiring process, assessment software can indeed reduce gender inequality in the labor market, by ensuring that hiring decisions are allocated meritocratically across candidates, thereby improving the overall qualifications of hires. This is because assessment software is designed to evaluate job applicants based on objective criteria, such as job applicants' skills, qualifications, and experience, rather than subjective factors, such as gender. Additionally, assessment software can be used to remove some of the human errors that can occur during the hiring process. For example, it can be designed to eliminate the possibility of recruiters subjectively interpreting resumes, which can lead to unconscious bias. This can result both in more diverse and higher qualified hires. Conversely, while predictive algorithms perform as good as assessment software in ensuring qualified applicants, they do not differ from human recruiters in the diversity of the selected applicant pool.

The most important concern raised by my analysis is the possibility of omitted variables, which can lead to biases in the results presented in this chapter. The algorithm bases its hiring choices on job applicants' characteristics I observe and may therefore miss unobservables that allow to accurately quantify human recruiters' bias in deciding whether a job applicant represents a successful hire, making my results an over or under estimation of the true effect of assessment software.

Chapter 4 draws on such a limitation of this chapter to develop a study about how assessment software and predictive algorithms heterogeneously affect diversity and quality of recruiters' hiring choices. In this chapter I asked whether assessment software and predictive algorithms can affect diversity and qualifications of new hires when they automate the hiring process. In chapter 4, I ask, instead, whether assessment software and predictive algorithms can affect hiring diversity and productivity when they augment human knowledge by providing information to recruiters. The approach in chapter 4 is complementary to this chapter in that it allows for selection on unobservables and does rely on modeling the human decision.

1.3.3 Chapter 4: AI and employers' choices (coauthored with Arthur Schram)

This chapter answers to RQ3a: 'How do assessment software and predictive algorithms heterogeneously affect gender inequality in the labor market when used together with human recruiters?' and RQ3b: 'What is the impact of assessment software's and predictive algorithms' evidence on human recruiters' prior beliefs?'

Laboratory, online, field, and audit experiments have been useful in uncovering mechanisms behind employers' discriminatory actions. Online experiments can be useful in research on gender because they allow for a larger and more diverse sample size than traditional in-person experiments. Additionally, online experiments can be less expensive and time-consuming to conduct. Further, online studies allow to reach participants from all over the world with various demographic characteristics, compared to in-person lab experiments, that are usually conducted within universities.

In this chapter, we conduct an online experiment, backed by a theoretical model, with 600 UK participants. We make participants act as recruiters. The experiment follows the conventional literature on information cascades, where participants develop their prior beliefs according to a private information they have on the value of two job candidates. As the experiment proceeds, participants update their prior beliefs according to the information we give them about the two job candidates. One such information comes from an assessment software and the other from predictive algorithms, which are available to random groups of participants and provide them with the actual individual productivity-value of each job candidate.

We show that both assessment software and predictive algorithms allow recruiters to escape information cascades. Additionally, both AI tools improve the overall productivity of selected applicants. Assessment software can also make employers change their prior beliefs about job candidates and enhance the diversity of hires, particularly when significant productivity differences exist among the job applicants under consideration.

This chapter answers to the central research question of the thesis by highlighting both theoretically and empirically that when it augments human knowledge in hiring, assessment software can make recruiters change their prior beliefs in the hiring process. Therefore, it has the potential to reduce gender inequality in the labor market by removing some of the unconscious biases that can influence hiring decisions. By providing employers with more objective and accurate information about job applicants, assessment software can help challenging and changing any prior beliefs or stereotypes that may influence employers' hiring decisions. For example, if employers are subject to the cultural stereotype that women are not as suitable for a particular job as men, and the assessment software provides strong evidence that contradicts this belief, employers' stereotype can be counteracted. However, the chapter suggests that in order for assessment software to ultimately challenge employers' prior beliefs, the contradictory evidence regarding job applicants' productivity should significantly favor women over men, allowing them to stand out within the applicant pool.

2

A new turning point for women

Abstract

This chapter examines whether firms' use of AI affects their proportion of female managers hired. The study uses panel data from the 500 largest companies in Europe and the US based on their revenues and employs a difference-in-differences approach. Despite concerns raised in the existing literature regarding AI being biased, the findings show that, on average, firms' adoption of AI leads to a 40% increase, relative to baseline, in the hiring of female managers. Further, firms' use of AI correlates to a reduction in firms being sued for gender discrimination in hiring.

2.1 Introduction

As discussed in chapter 1, firms employ AI in the hiring process with the aim to improve efficiency and accuracy. Their belief is that hiring algorithms can help alleviate biases among recruiters and introduce objectivity (Langenkamp et al., 2020). However, existing research indicates that AI can, in fact, exhibit gender-based discrimination, thereby raising doubts about its objectivity (Cowgill and Tucker, 2020; Gebru, 2020; Cowgill, 2019; O'Neil, 2016).

The question arises as to how we can explain the paradox of a technology designed to eliminate human bias in decision-making actually being biased itself. Chapter 1 sheds light on this issue by exploring the effect of two types of AI tools used in hiring: predictive algorithms and assessment software. The answer to this question, I argue, lies in the specific type of AI employed by firms in the hiring process. As outlined in chapter 1, predictive algorithms tend to perpetuate existing gender inequalities in the labor market (Cowgill and Tucker, 2020; Gebru, 2020; Daugherty, Wilson, and Chowdhury, 2019; Cowgill, 2019; O'Neil, 2016). Conversely, when firms employ assessment software, they have the potential to reduce gender inequality in the labor market by basing hiring decisions solely on job applicants' performance rather than on ascribed characteristics such as gender (Daugherty, Wilson, and Chowdhury, 2019; Silberg and Manyika, 2019).

While these studies on assessment software suggest that it has the potential to mitigate gender inequality in the labor market, they do not offer empirical evidence supporting such an effect. In this chapter, I empirically contribute to the existing literature by setting the ground of an in depth analysis about firms' use of AI in hiring by examining the combined effect of assessment software and predictive algorithms in hiring on gender inequality in the labor market. The primary focus is on addressing the following research question: what is the effect of AI on gender inequality in the labor market? To explore this question, I build upon the theoretical framework introduced in chapter 1 and develop hypotheses regarding the influence of AI on gender inequality in employment outcomes. These hypotheses are empirically tested using a staggered difference-in-differences approach on panel data encompassing the 500 largest companies in Europe and the US over an 8-year period

(2013-2021).

To measure gender inequality in employment outcomes, I specifically analyze gender disparities within companies' managerial pools. Gender inequality in leadership positions holds significant economic consequences for women, as it accounts for a significant portion of the gender wage gap in the labor market (Mandell et al., 2022).

The findings show that, on average, firms' use of AI leads to a 1.5 percentage points increase, an effect of 40% at baseline, in the hiring of female managers. Further analysis reveals that the use of AI also correlates with a reduction in gender discrimination lawsuits related to hiring practices.

2.2 Theoretical framework

When individuals apply for managerial positions at a company, they bring their unique set of skills.

When a company decides to hire someone as a manager, its primary goal is to maximize its profit. The company's profit is closely linked to the performance and skills of the manager it hires.

The company assesses job applicants and has its own method for selecting highly productive candidates based on certain probabilities. This probability is determined by the information available in job applicants' resumes, interviews, or other sources, such as employer's or employees' informal networks (Fernandez et al., 2000).

Now, consider that job applicants can be either male or female. Previous research indicates that when the applicant pool consists of individuals of different genders, discrimination can occur due to employers holding biased beliefs about the competence of women and men (Ridgeway, 2011, 2001). Since recruiters do not have perfect information about the future performance of job applicants they might hire, they may rely on easily observable characteristics like gender when making hiring decisions. As a result, employers tend to discriminate against women due to these biased beliefs about the competence of female and male job applicants. These biased beliefs, known as "status beliefs," stem from widely held cultural beliefs that associate greater social significance and overall competence, as well as specific skills, with one gender category (men) over the other (women) (Ridgeway, 2001, p. 638). Such cultural beliefs about gender can contribute to gender discrimination in hiring by distorting recruiters' expectations of the unique productivity of men and women (Ridgeway, 2011; Heilman and Okimoto, 2007; Gorman, 2005). Furthermore, the cost of obtaining information about individual job applicants' productivity often leads recruiters to rely on group statistics (average male and female performance in the labor market) to estimate candidates' unique productivity (Arrow, 1973; Phelps, 1972). As a result of either statistical or status discrimination, male job applicants tend to be selected with a higher probability than female job applicants, even when male and female candidates possess the same skills.

Now, consider that companies can adopt AI algorithms to reduce the cost of evaluating job applicants. As we know from existing research and from chapter 1, AI algorithms offer several advantages for firms: they provide abundant information about potential employees, and they process this information more quickly and with less effort compared to human recruiters.

In this chapter, my interest lies in evaluating the impact of AI adoption by companies on the proportion of female managers hired, without distinguishing between the types of AI used. This is because there is a lack of research on the effects of AI on managerial hiring. Furthermore, since companies can use assessment software or predictive algorithms in their hiring processes, understanding the combined impact of these two approaches is crucial for laying the groundwork for a comprehensive analysis of the various effects of these AI tools, which I will conduct in Chapter 3.

From Chapter 1, we know that predictive algorithms utilize historical data to estimate a candidate's likelihood of success based on the past performance of employees. However, if this historical data contains biases, the algorithms may perpetuate gender inequality in hiring (Gonzalez et al., 2022; Köchling and Wehner, 2020; Cowgill and Tucker, 2020). We also know that assessment software evaluates candidates based on their performance in cognitive tests, interviews, and other assessments, providing accurate information about individual candidates and reducing the influence of biased beliefs and decisions. Therefore, it is likely to decrease gender inequality in hiring (Williams, 2022; Raisch and Krakowski, 2021; Black and van Esch, 2020; Daugherty et al., 2019). The combined impact of these two approaches may lead to either a reduction in gender discrimination if firms use more assessment software than predictive algorithms, or an increase in gender discrimination if firms use more predictive algorithms than assessment software. Hence, the empirical analyses of this chapter are grounded in two conflicting hypotheses:

Hypothesis 2.2.1 *The use of AI in hiring has the potential to reproduce existing gender inequality in firms' managerial roles due to the influence of historical biases in the data used for AI training.*

Hypothesis 2.2.2 *The use of AI in hiring has the potential to reduce existing gender inequality in firms' managerial roles due to the avoidance of relying on historically biased data.*

2.3 Data and empirical strategy

2.3.1 Data

The analysis conducted in this chapter is based on European and American firms that were listed in the Fortune Global 500 in 2021. This dataset comprises European and American companies that rank among the top 500 corporations globally in terms of revenue. The decision to focus on these specific firms is primarily driven by the availability of publicly accessible information regarding their utilization of AI in the hiring process. Additionally, these firms exhibit financial and organizational similarities. Specifically, they present comparable levels of productivity and size, as depicted in Table 2.1. These factors, as highlighted by Gòmez and Vargas (2012), influence their innovative capabilities and, consequently, their likelihood of adopting AI in the hiring process.

The availability of publicly accessible data on AI usage and the similarity among companies enable me to investigate the impact of firms' adoption of AI on their proportion of female managers hired. To conduct this analysis, I employ a matching approach, pairing

Table 2.1: Summary statistics

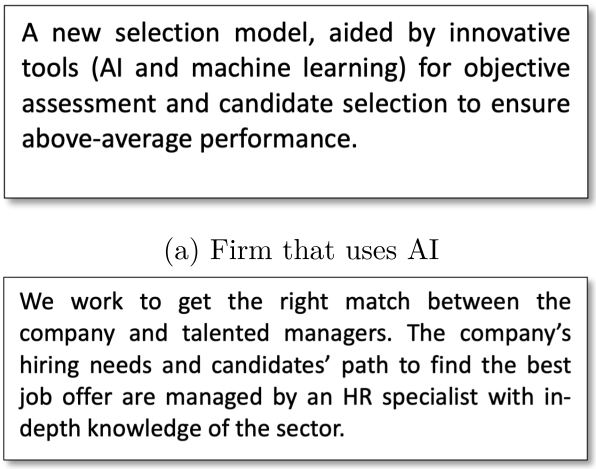
	Full sample		Restricted matched sample		Mean difference
	Mean	SD	Mean	SD	
Share of female managers hired	0.026	0.05	0.024	0.042	0.002
Use of AI	0.107	0.309	0.298	0.458	0.191*
Log assets	25.262	1.469	25.748	1.368	0.486*
Log productivity	24.406	0.824	24.349	0.528	0.057
Net income (millions)	12.357	18.577	12.629	13.382	0.272
Profit margin	13.286	14.108	15.353	11.996	2.067*
Return on equity	27.986	67.858	16.954	12.365	11.032*
Log employees	11.278	1.115	11.489	0.63	0.211*
Share of female directors	0.052	0.044	0.053	0.039	0.001
N	1,621		531		

Significance of difference in means: *** p<0.01, ** p<0.05, * p<0.1

firms that utilize AI with the most comparable firms that do not employ it. Firms' characteristics data are sourced from the Orbis database by Bureau van Dijk[1]. I gather information on variables that could potentially influence differences in the adoption of new technologies among firms, such as the firm's headquarters location, industry, productivity, total assets, return on equity, profit margin, net income, number of employees (Antonelli, Orsatti, and Pialli, 2022; Gòmez and Vargas, 2012), and the proportion of female directors as a control for gender diversity in leadership roles within firms. It's important to note that the share of female directors differs from the share of female managers since directors are nominated by the firm's shareholders, while managers are hired. Therefore, AI would not directly impact gender (in)equality within the board of directors. Data spanning from 2013 to 2021 is collected for the analysis.

The dependent variable in this study is the proportion of female managers within firms. This variable represents the annual number of female managers hired by each firm relative to the total number of managers hired in that year. Conversely, obtaining data on firms' use of AI in hiring is more challenging. To determine whether firms utilize AI tools in their hiring processes and the timing of their adoption, I manually extract information on AI usage from firms' publicly available annual integrated reports. Figure 2.1 provides an example of the information regarding AI use in hiring found in a firm's annual integrated report. Panel a shows an extract of the annual integrated report of a company in my data, which uses AI in hiring. Panel b shows an extract of the annual integrated report of a company in my data, which does not use AI in hiring.

[1] https://www.bvdinfo.com/en-gb/our-products/data/international/orbis



A new selection model, aided by innovative tools (AI and machine learning) for objective assessment and candidate selection to ensure above-average performance.

(a) Firm that uses AI

We work to get the right match between the company and talented managers. The company's hiring needs and candidates' path to find the best job offer are managed by an HR specialist with in-depth knowledge of the sector.

(b) Firm that does not use AI

Figure 2.1: Example of the information regarding AI use in hiring in a firm's annual integrated report

I assign each firm to the AI treatment group ($D_i = 1$) if its annual integrated report provides evidence of using AI tools for personnel decisions or evaluating job candidates.

2.3.2 Empirical strategy

The empirical strategy employed in this study employs a staggered difference-in-differences approach with ex-ante matching. The decision to perform ex-ante matching is motivated by the larger number of firms in the control group compared to the treatment group. By conducting matching, potential outliers can be eliminated, and the parallel trend assumption of the difference-in-differences specification can be strengthened.

The identification strategy is structured in two stages:

(i) In the first stage, I match treated and control firms based on observable characteristics using Mahalanobis distance matching (Mahalanobis, 1936). This matching technique aims to create comparable pairs of firms.

(ii) In the second stage, a staggered difference-in-differences estimation is performed with multiple time periods. This approach follows the methodology proposed by Callaway and Sant'Anna (2021). The reason for using this approach is that not all treated firms in my data adopted AI in hiring during the same year, necessitating the consideration of multiple time periods to capture the staggered nature of the adoption process.

Mahalanobis distance matching

The first step of the empirical strategy involves using Mahalanobis distance matching to achieve balance between the treated and control firms in terms of observable covariates measured prior to the year of AI adoption by the treated firms. By employing Mahalanobis distance matching before implementing the difference-in-differences approach, we can account for the systematic dynamic differences between firms that use AI in hiring and those that do not.

For each treated firm, where treatment D_i denotes the use of AI in hiring, I identify all available untreated firms with the most similar variables x based on the Mahalanobis distance metric. These variables are selected based on previous research indicating their relevance in determining firms' adoption of AI in hiring. Specifically, firm size, including total assets and number of employees, is known to significantly influence the probability of technology adoption (Gòmez and Vargas, 2012). Therefore, I match treated and control firms based on measures of firm size. Additionally, the likelihood of technology adoption is strongly associated with firms' productivity, profit margins, and return on equity, so I also match firms based on these variables. Considering that firms may adopt AI to improve diversity, I match firms based on the share of women in the board of directors. Finally, as the relationship between innovation and the aforementioned variables varies across countries and industries (Damanpour, 1992), I match treated and control firms based on industry (using the 4-digit NACE code) and country.

By carefully matching the treated and control firms on these relevant covariates, I aim to ensure comparability and minimize potential biases when evaluating the impact of AI adoption on hiring outcomes. Equation 2.1 presents the econometric specification of the Mahalanobis distance definition.

$$d(u, v) = (u - v)TC - 1(u - v) \tag{2.1}$$

With u and v values of $\{xT, q^(x)\}T$, where x are the observable covariates and $q^(x)$ is the estimated log odds against exposure to treatment; and C sample covariance matrix of $\{xT, q^(x)\}$ in the control group (Rosenbaum and Rubin, 1985).

The Appendix reports evidence regarding the balance of the observable covariates among treated and control firms, before and after matching. In particular, Figure 2.6 presents the balancing achieved after the Mahalanobis distance matching.

Staggered difference-in-differences

I estimate the impact of using AI on the proportion of female managers hired by firms through a staggered difference-in-differences approach with multiple time periods. This estimation method is based on the approach proposed by Callaway and Sant'Anna (2021).

Estimating the ATT

The core of the analysis relies on calculating the average treatment effect on the treated (ATT) of AI adoption on the share of female managers hired. In this context, a firm is defined as ”treated” based on its year of AI adoption. Specifically, I aim to estimate the group-time average treatment effect for firms belonging to each treated group (denoted as g representing the year of AI adoption for each firm) at each time period (denoted as t) This estimation is formalized in equation 2.2, following the framework outlined in Callaway and Sant'Anna (2021), where Y_t denotes the share of female managers hired by firms in each year ”t”.

$$ATT(g, t) = E[Y_t(g) - Y_t(0)|G_g = 1] \tag{2.2}$$

In this equation, I compare the outcomes of firms that adopted AI in year ”g” with those that did not adopt AI (i.e., $G_g = 0$) in the same year. The difference in outcomes captures the causal effect of AI adoption on the proportion of female managers hired by firms.

Accounting for Observable Covariates

To enhance the precision and validity of my ATT estimates, I apply a weighted approach. Each observation is weighted using Mahalanobis distance matching, aligning the estimate of ATT(g, t) with observable covariates. This step helps to account for any potential pre-treatment confounding factors and ensures that my estimates are robust and credible.

Threats to the identification

In this section, I discuss several critical assumptions that underpin the identification strategy employed in this study. These assumptions are central to the reliability of my ATT estimates (Callaway and Sant'Anna, 2021) and warrant careful consideration.

Limited treatment anticipation posits that firms should not anticipate treatment by any period. While this assumption serves as a cornerstone of my analysis, it is essential to acknowledge its potential vulnerabilities. Specifically, I cannot definitively assert that firms do not increase the hiring of female managers in anticipation of AI adoption. The introduction of AI may be part of a broader effort to enhance gender diversity, thus potentially influencing firm behavior before the actual treatment. To address this concern, I account for potential anticipatory behavior by allowing for such behavior and imposing conditional parallel trends in pre-treatment periods, thereby making the parallel trend assumption discussed in the subsequent paragraph stronger.

Conditional parallel trends rely on a comparison between treated firms and a never-treated group. As advocated by Callaway and Sant'Anna (2021), rather than comparing treated firms with not-yet-treated ones, I compare treated firms with never-treated ones, because I have a sizable group of firms that abstain from participating in the treatment throughout the observation period, constituting approximately 70% of the firms in my restricted matched sample. This assumption postulates that, conditional on covariates, the average outcomes for firms initially treated in group g (with g representing the year of AI adoption) and the never-treated firms would have followed parallel paths in the absence of the treatment. The results section subsequently presents evidence affirming the validity of this conditional parallel trend assumption.

Inversibility of the treatment, also known as staggered treatment adoption in the literature, asserts that no unit is treated at time t = 1, and once a unit becomes treated, it remains treated in the subsequent periods. To visually depict this assumption, Table

2.2 and Figure 2.2 illustrate the yearly share and number of firms entering the treatment status and persisting as treated entities until the final year of observation. Moreover, Table 2.2 and Figure 2.2 confirm that no units are treated at time t = 1 (2013), thereby substantiating the adherence of my data to the assumption of treatment inversibility.

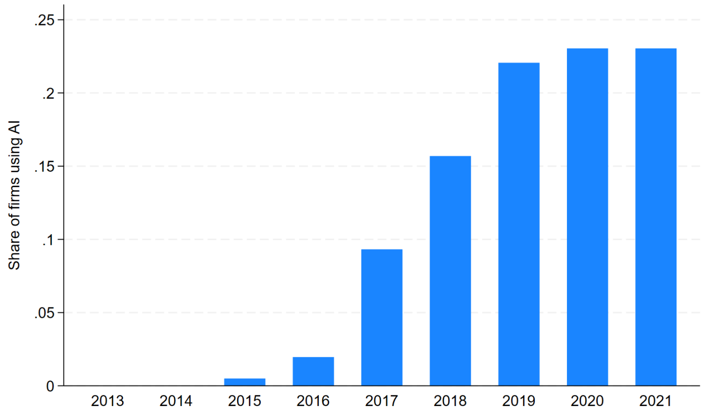


Figure 2.2: Firms entering the treatment status and remaining treated

Table 2.2: Firms entering the treatment status and remaining treated

Year	N
2013	0
2014	0
2015	1
2016	4
2017	19
2018	32
2019	45
2020	47
2021	47

2.4 Results

2.4.1 Effect of AI use on firms' share of female managers hired

This section presents the average treatment effect for the treated (ATT (g, t)) of firms' use of AI on their share of female managers hired. Figure 2.3 shows the graphical representation of the standardized ATT (g, t) estimate. Table 2.3 reports the event study and simple weighted average standardized estimates of the ATT (g, t). Standard errors are clustered at firm level.

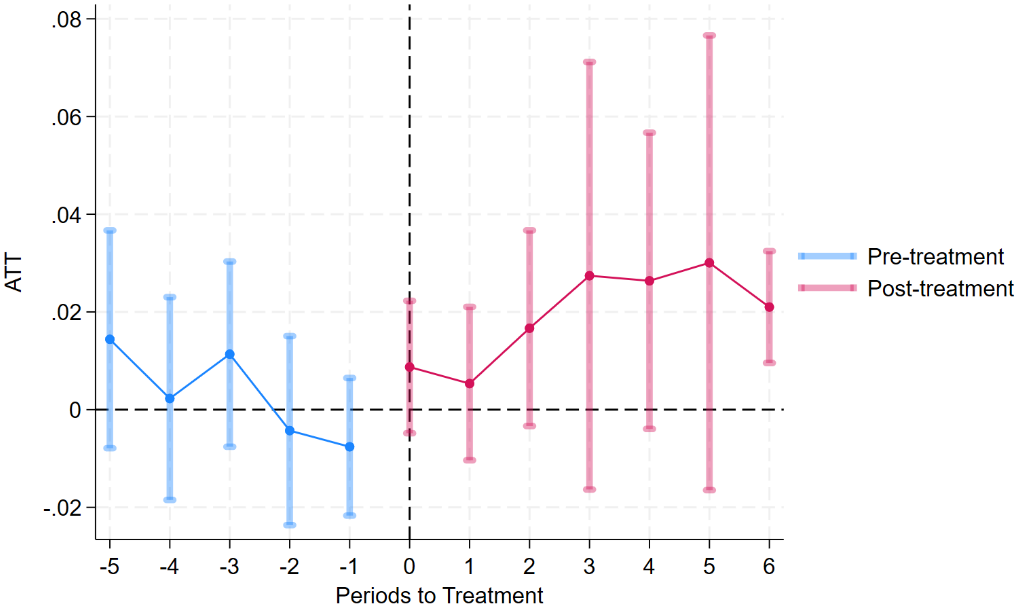


Figure 2.3: Effect of AI use on the share of female managers hired. Event study

Table 2.3: Effect of AI use on the share of female managers hired

	Share of female managers hired (1)
Use of AI in hiring	0.015** (0.006)

Standard errors in parentheses, clustered at firm level
*** p<0.01, ** p<0.05, * p<0.1

NOTES: The table reports the ATT for firms matched with Mahalanobis distance matching

As depicted in figure 2.3, the conditional parallel trend assumption is satisfied since the pre-treatment average estimates of the ATT and the estimates of the ATT in each year before treatment ($t-5$ to t_0) are not statistically different from zero. The event study analysis suggests that, if any, the pre-treatment trend would be negative in magnitude.

The estimated simple weighted average ATT in table 2.3 indicates that firms' use of AI in the hiring process leads to an average increase of approximately 1.5 percentage points in the share of female managers hired. This effect corresponds to a 40% increase, relative to baseline. The Appendix shows a placebo test with fictious dates of AI adoption.

The findings of the estimation suggest that the use of AI by firms can contribute to reducing the persistent underrepresentation of women in managerial roles. An important implication is that AI adoption in hiring can promote gender equality in the labor market by increasing female representation in managerial positions.

2.4.2 AI and gender discrimination

The main finding of this study, which reveals that the use of AI by firms increases the proportion of female managers hired, can potentially be attributed to the AI's ability to reduce gender discrimination in the hiring process. Unfortunately, the available data in this paper does not allow for a direct test of this mechanism. However, I can conduct an exploratory analysis to examine whether there is any correlation between firms' use of AI and gender discrimination lawsuits in hiring.

To conduct this analysis, I gathered publicly available information on court cases involving firms that were sued for gender discrimination in hiring between 2013 and 2021. Due to the accessibility of lawsuit data and documentation, I was only able to collect information on firms in the United States. Consequently, the results of the following analyses do not apply to both European and US firms. Nonetheless, as demonstrated in Figure 2.4 and Table 2.4, limiting the analysis to US firms does not compromise the reliability of the findings. In fact, the effect of AI usage on the proportion of female managers hired in US firms closely aligns with the estimated ATT for both European and US firms. Therefore, I can narrow down the sample to solely US firms, which accounts for 51% of the original sample size.

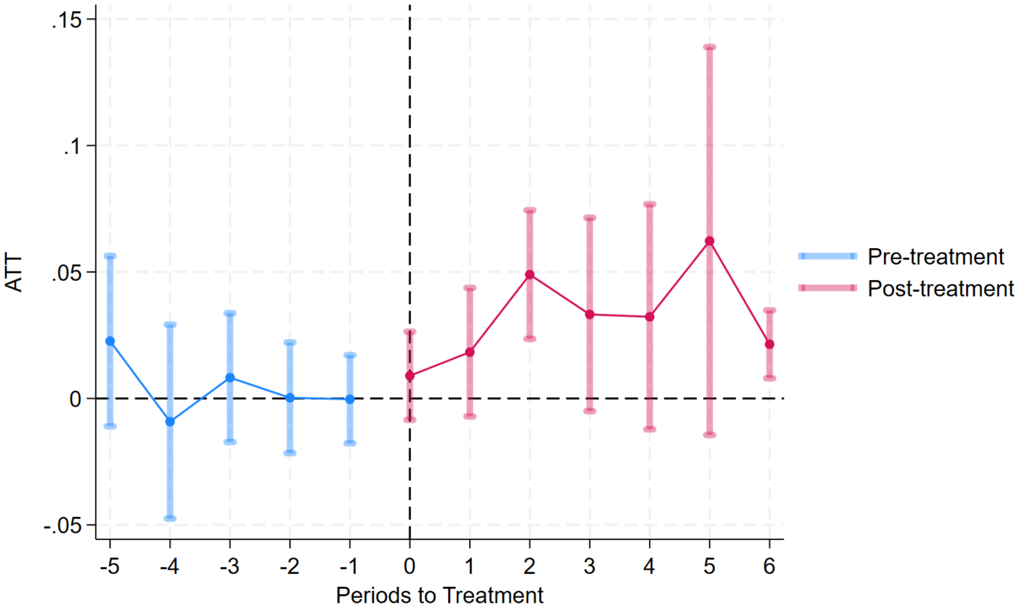


Figure 2.4: Effect of AI use on the share of female managers hired. Event study. US firms

Table 2.4: Effect of AI use on the share of female managers hired. Weighted ATT. US firms

	Share of female managers hired
	(1)
Use of AI in hiring	0.019** (0.008)

Standard errors in parentheses, clustered at firm level
*** p<0.01, ** p<0.05, * p<0.1

NOTES: The table reports the ATT for firms matched with Mahalanobis distance matching

Figure 2.5 shows the graphical results of the effect of using AI on firms' probability of being sued for gender discrimination in hiring for managerial jobs. Table 2.5 reports the event study and simple weighted average standardized estimates of the effect of using AI on firms' probability of being sued for gender discrimination in hiring for managerial jobs.

Table 2.5: Effect of AI use on the probability of being sued for gender discrimination in hiring for managerial jobs. Weighted ATT. US firms

	Probability of being sued for gender discrimination in hiring managers
	(1)
Use of AI in hiring	-0.15* (0.088)

Standard errors in parentheses, clustered at firm level
*** p<0.01, ** p<0.05, * p<0.1

NOTES: The table reports the ATT for firms matched with Mahalanobis distance matching

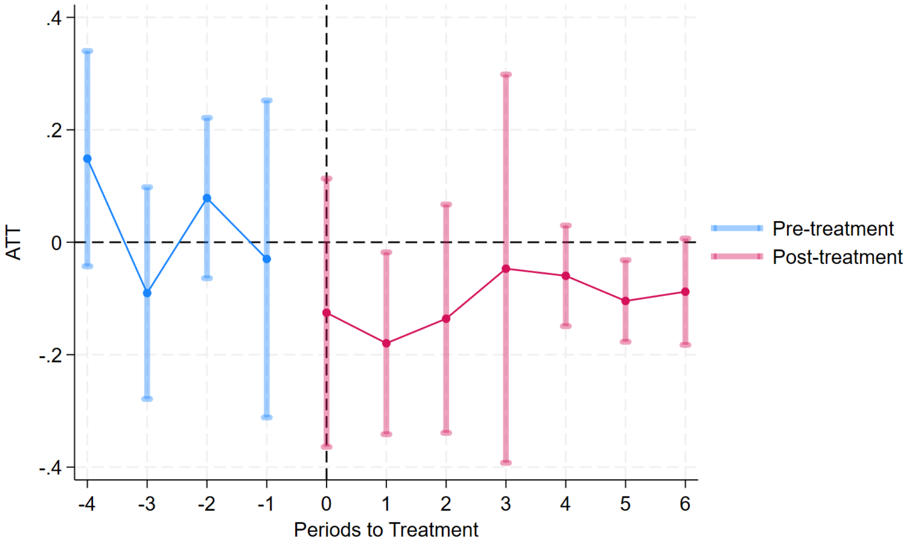


Figure 2.5: Effect of AI use on the probability of being sued for gender discrimination in hiring for managerial jobs. Event study. US firms

Using AI is associated with a 15 percentage points decrease in the likelihood of firms being sued for gender discrimination in hiring. This finding suggests that the adoption of AI may assist firms in reducing gender discrimination during the hiring process, ultimately leading to an increase in the proportion of female managers hired.

2.5 Conclusion

This chapter is motivated by the need to examine the impact of AI on gender inequality in hiring. While firms adopt AI with the expectation that it will help mitigate biases and promote gender equality in hiring decisions, previous research suggests that AI may actually perpetuate gender-based discrimination.

Using a staggered difference-in-differences approach, this chapter reveals that firms' use of AI in hiring decreases gender inequality in hiring outcomes and is associated with a reduced likelihood of firms facing lawsuits for gender discrimination in hiring.

This chapter provides new evidence on the relationship between AI and gender inequality in hiring. Although the results in this chapter show that AI can enhance the hiring of women for managerial positions, the specific type of AI driving this effect remains unknown. Chapter 3 addresses this limitation by comparing the two tools with human recruiters and understanding the extent to which each type of software may reduce or increase gender inequality in hiring.

Further, the measurement of AI usage in hiring relies on self-reported micro-level data obtained from firms. While addressing potential endogeneity concerns, this data allows for estimating the overall effect of assessment software and predictive algorithms on gender inequality in hiring outcomes by assessing their influence on the hiring process. Nevertheless, information on whether firms use these hiring tools for automating decisions, augmenting human knowledge, or both is not available.

To address this limitation, chapter 3 focuses on evaluating whether assessment software and predictive algorithms can reduce gender inequality in the labor market when they are used to automate the hiring process, while chapter 4 examines their impact when used to enhance human knowledge in hiring.

Appendix

2.5.1 Balance of observable covariates

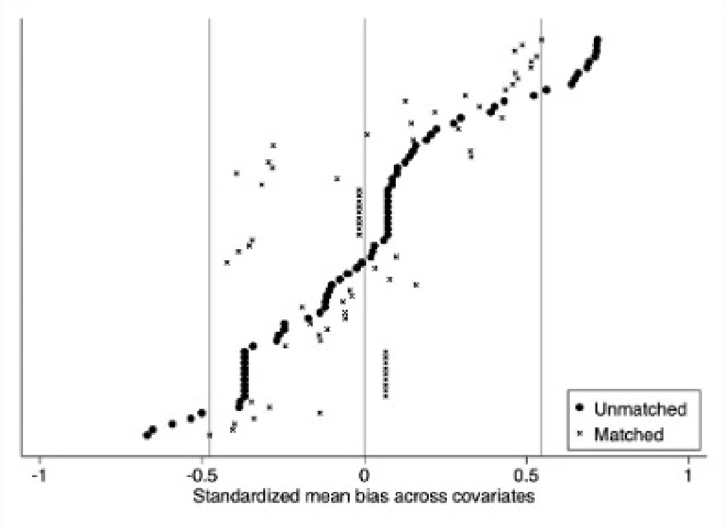


Figure 2.6: Balancing of the observable covariates

Figure 2.6 shows the standardized mean bias for all covariates before and after matching, that is the difference of the means in the treated and non-treated firms as a percentage of the square root of the average of the sample variances in the treated and non-treated groups (Rosenbaum and Rubin, 1985). The mean absolute standardized bias across covariates after matching is 21.6, which is smaller than the absolute standardized mean bias across covariates before matching (29). As Figure 4 shows, matching reduced the standardized mean bias to less than ∼0.5 for all covariates. Refinement is desirable, but matching has done well at balancing the treated firms and their control counterparts, adjusting reliably for all the covariates.

2.5.2 Placebo test with fictious dates of AI adoption

Figure 2.7 and table 2.6 show the estimated Average Treatment Effect (ATT) of firms' use of AI on their share of female managers hired after randomly assigning the year of AI adoption to treated firms.

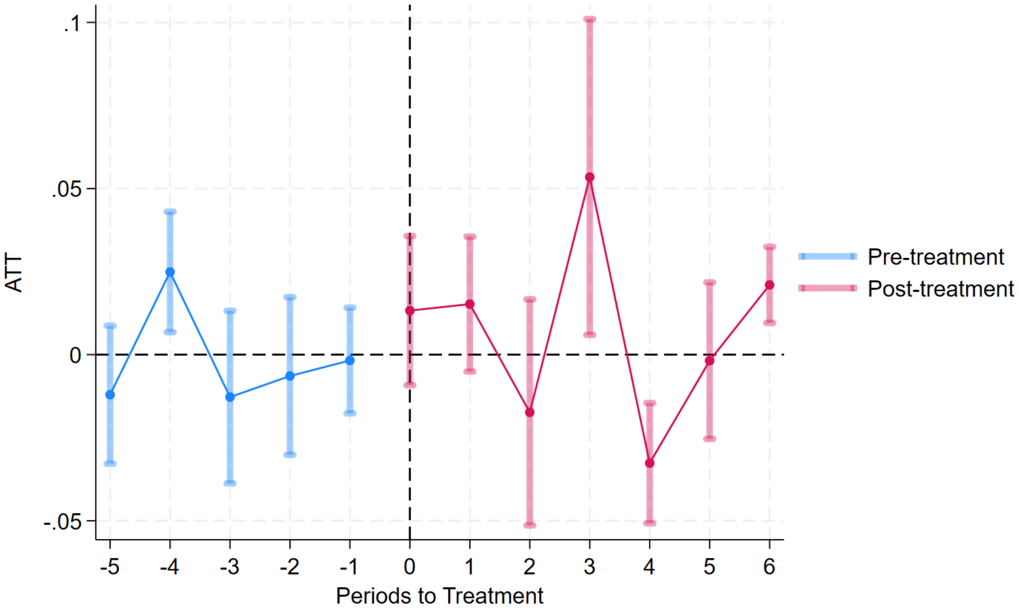


Figure 2.7: Effect of AI use on the share of female managers hired. Event study

Table 2.6: Effect of AI use on the share of female managers hired

	Share of female managers hired (1)
Use of AI in hiring	0.015 (0.009)

Standard errors in parentheses, clustered at firm level
*** p<0.01, ** p<0.05, * p<0.1

NOTES: The table reports the ATT for firms matched with Mahalanobis distance matching

Figure 2.8 and table 2.7 show the estimated Average Treatment Effect (ATT) of firms' use of AI on their share of female managers hired after randomly assigning firms to the treatment and control groups and to a random year of AI adoption.

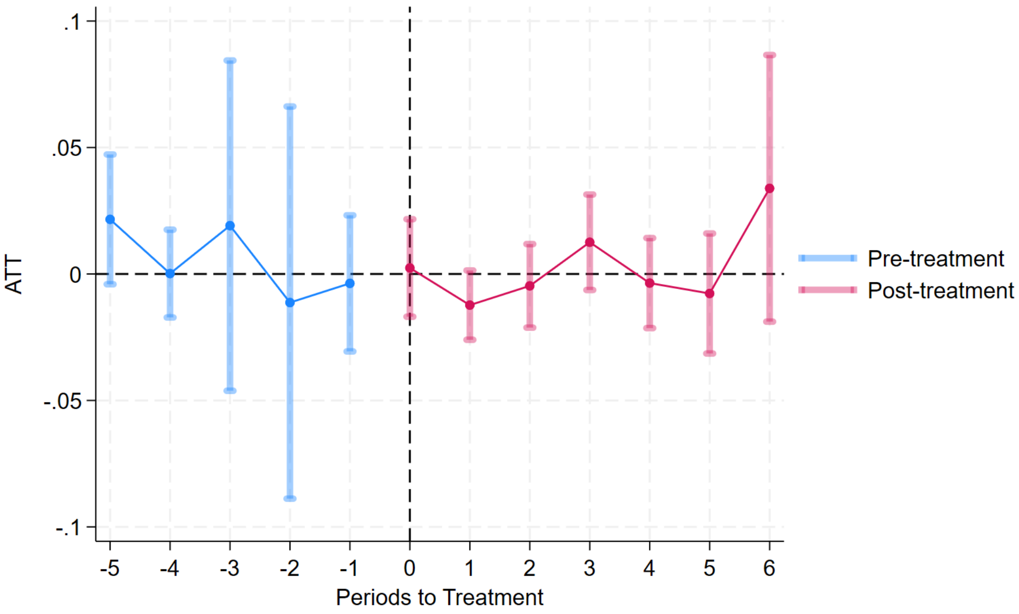


Figure 2.8: Effect of AI use on the share of female managers hired. Event study

Table 2.7: Effect of AI use on the share of female managers hired

	Share of female managers hired
	(1)
Use of AI in hiring	0.006 (0.007)

Standard errors in parentheses, clustered at firm level
*** p<0.01, ** p<0.05, * p<0.1

NOTES: The table reports the ATT for firms matched with Mahalanobis distance matching

3

The hiring dilemma: efficiency, equality, or both?[1]

Abstract

This chapter examines hiring as a balance between efficiency and equality. When hiring, firms aim to minimize the cost of screening and selecting job applicants while also striving for gender equality. Indeed, modern hiring algorithms are designed to reduce costs. However, existing research shows that AI technologies based on predictive algorithms may be biased against women. This chapter investigates the impact of predictive algorithms and assessment software on gender inequality when the tools fully automate hiring. I analyze data from a private firm's recruitment process and develop two resume screening algorithms — one that functions as a predictive algorithm and another as an assessment software. In particular, one resume screening algorithm evaluates job applicants by assessing their skills, while the other predicts the best applicant to hire based on incumbent workers' skills and characteristics. I show that, while the predictive algorithm does not differ in terms of gender equality and qualifications of the selected applicant pool compared to human recruiters, the assessment software not only doubles the probability of selecting female applicants compared to both human recruiters and predictive algorithms but also significantly improves the qualifications level of the selected candidates.

[1] The intervention study described in this chapter and the resulting dataset have been used in another shorter and different paper of mine, which is published: Pisanelli, E. (2022). Your resume is your gatekeeper: Automated resume screening as a strategy to reduce gender gaps in hiring. *Economics Letters*, 221:110892.

3.1 Introduction

In Chapter 2, we gained insights into the impact of AI on gender inequality in hiring, specifically focusing on the combined effect of predictive algorithms and assessment software. I found that AI has the potential to help firms reduce gender inequality in the labor market. However, the analysis in chapter 2 lacks a heterogeneity analysis that shows which AI tool drives such an effect.

Building upon these findings, this chapter delves deeper into the relationship between AI and gender inequality in hiring, with a specific focus on comparing the effect of assessment software and predictive algorithms to human recruiters. Also, while chapter 2 provided valuable insights into whether AI can contribute to reducing gender inequality in hiring outcomes, it remains unclear whether this holds true when predictive algorithms and assessment software are granted full autonomy in the hiring process or when they are used in conjunction with human recruiters. Therefore, the objective of this chapter is to explore the inner workings of assessment software and predictive algorithms and examine how they influence gender inequality in hiring when they automate the hiring process. Specifically, the chapter answers to the following research questions: How do assessment software and predictive algorithms heterogeneously affect the qualifications level and diversity of hires when they automate the hiring process (*i.e.*, replace human decision-making)? Do assessment software, predictive algorithms and human recruiters differ in this respect?

This chapter is the first to conduct an intervention study within a private firm to examine how assessment software and predictive algorithms affect the diversity and quality of the firm's hiring process. Assessment software and predictive algorithms select successful job applicants based on different strategies, i.e., respectively, by relying on objective and standardized assessments of job applicants' skills (Li et al., 2021) or by using the characteristics and success of previous employees (Rhea et al., 2022). I, thus, expect (also in light of Chapter 2) that when these algorithms are used to automate the hiring process, they will have different effects on the firm's hiring outcomes compared to each other.

The approach taken in this chapter starts with the understanding that the hiring process involves a balance between efficiency and equality, with equality referring to the meritocratic allocation of rewards (Jackson, 1998). Firms seek the best applicant while attempting to minimize the costs of the hiring process. Recently, firms have also been increasingly adopting merit-based practices to enhance equality and linking hiring, rewards, and promotion decisions solely to performance rather than ascribed characteristics such as gender (Castilla, 2008). How does this practice translate into AI-powered hiring? While both predictive algorithms and assessment software can reduce firms' hiring costs (Sharma, 2018), chapter 1 shows that assessment software has the potential to increase diversity in firms' hiring processes, while predictive algorithms are likely to reproduce existing human biases. The use of assessment software may enable firms to reduce hiring costs while simultaneously enhancing the gender equality of selected applicants. This is particularly achievable if the software is designed to be gender-blind and eliminates human discretion from hiring decisions. In contrast, predictive algorithms should not significantly differ from human recruiters (Cheng and Hackett, 2021).

This chapter empirically analyzes the decision to interview job candidates based on resume screening for an international sales and purchasing agent position. Initial interviews decisions based on resume screening are crucial in determining the gender composition of the new hires (Fernandez-Mateo and Fernandez, 2016; Fernandez and Fernandez-Mateo, 2006; Petersen and Saporta, 2004). As existing research shows, recruiters bend the pipeline against women in the resume screening phase of hiring, after which all qualified candidates, regardless of gender, have equal chances to be hired (c.f. Fernandez-Mateo and Fernandez (2016) for a discussion about the gender gap before and after the resume screening stage of hiring). It is, thus, fundamental to study how assessment software and predictive algorithms affects both the qualifications level and diversity of the applicant pool shortlisted after resume screening. The international sales and purchasing agent is classified according to ISCO (International Classification of Occupations) under code 3, which refers to high-skill jobs. The firm providing the data for this study has a history of predominantly male workers in this role, indicating a lack of gender diversity. The data consists of information from resumes submitted by candidates for the position. Like many other medium-sized firms, the firm in question receives numerous applications for each vacancy and rejects the majority of candidates based on an initial resume screening. The objective of this chapter is to answer the question of whether and how assessment software and predictive algorithms, which automate the hiring process, affect gender diversity and qualifications of new hires. To do so, I construct two resume screening algorithms that closely resemble the most commonly used resume screening algorithms available in the IT market. This was done based on the suggestions provided by a startup that specializes in developing hiring algorithms for companies. To ensure the reliability of my empirical strategy, it is important to note that the firm under study does not utilize AI in its hiring process and has never employed it prior to the commencement of this study.

By comparing the candidates selected by the algorithms to each other and to the actual selection decisions made by human recruiters within the firm, I present two key findings. First, while predictive algorithms do not differ from human recruiters in selecting a gender diverse pool of applicant, assessment software significantly increase the probability to select female applicants. Second, both assessment software and predictive algorithms increase the probability to select qualified candidates to 1.

These results illustrate that current hiring practices in firms are not efficient and leave room for algorithms to improve both diversity and quality of hiring. While it is easy to believe that both algorithms can ensure highly qualified applicants, even if assessment software shows the potential to significantly increase female representation in applicant pools, I remain cautious about concluding that assessment software inherently enhances gender diversity. In my specific setting, gender equality in job advertisements is not considered, and the way firms target or frame job ads may hinder the supply of female candidates, making it costly for gender equality even when assessment software is in place (Datta et al., 2018; Lambrecht and Tucker, 2019).

3.2 Theoretical framework

Based on the theoretical framework outlined in chapter 1, we know that when firms decide whether to hire a candidate, they incur a cost related to the hiring process, regardless of the outcome. Therefore, after hiring a new worker, the firm's overall profit is determined by the performance of the newly hired worker.

Before making the hiring decision, the firm reviews candidates' resumes and gathers information about job applicants. The firm's hiring decision depends on this information. Essentially, the firm aims to make a hiring decision that maximizes its profit from hiring the worker based on the available information about job applicants. Existing literature suggests that when firms have more accurate information about individual job candidates' performance, they are less likely to rely on heuristics and ascribed characteristics, such as gender, in their hiring decisions (Arrow, 1973; Phelps, 1972). Therefore, as the quality and impartiality of the information the firm obtains improve, the likelihood increases that the firm will make a decision to hire a high-performing candidate without considering gender.

This framework sets the stage for three different scenarios that the firm might encounter when evaluating information on job applicants for hiring purposes.

3.2.1 Firms do not use AI

Employers often face imperfect information regarding the future productivity of job candidates, which creates an incentive for them to rely on easily observable ascriptive characteristics, such as gender, to infer the expected productivity of applicants (Correll and Benard, 2006). As discussed in chapter 1, due to the high costs associated with acquiring individual workers' productivity information, employers often resort to using statistical information about groups (e.g., men and women) to make inferences about individual workers' productivity (Arrow, 1973; Phelps, 1972). Additionally, employers may rely on biased beliefs about the job candidate's group affiliation (e.g., men and women), which can stem from widely held cultural beliefs that associate greater social significance, general competence, and specific skills with one social category over another (Ridgeway, 2001, p. 638). This reliance on surrogate group-level information can contribute to gender discrimination in hiring (for a comprehensive review, see Correll and Benard, 2006).

The use of surrogate group-level statistical information and biased beliefs by employers during the hiring process does not accurately reflect job candidates' individual performance but instead triggers employers' gender stereotypes (Correll and Benard, 2006). Consequently, as indicated in previous research, firms may engage in gender discrimination when hiring new workers due to the influence of gender stereotypes on employers' decision-making (Correll, 2001; Ridgeway, 2001; Ridgeway, 2004; Correll and Benard, 2006; Ridgeway, 2011).

3.2.2 Firms use predictive algorithms

We already established in chapter 1 that predictive algorithms use firms' historical data on workers' productivity and characteristics as a benchmark for assessing the performance of job applicants (Rhea et al., 2022). Up until now, predictive algorithms have been designed to estimate the average probability of success for a specific group (e.g., men, women) based on the past performance of firms' employees. This estimation is then used to predict the future performance of individual job applicants belonging to the same group (Langenkamp, Costa, and Cheung, 2019). Hence, it is reasonable to argue that the functioning of predictive algorithms is not fundamentally different from the decision-making process of human recruiters. Moreover, predictive algorithms, by relying on data concerning firms' previous hires, link gender to the probability of success in a given job. If companies have predominantly hired men in the past, predictive algorithms inherit the biases and biased beliefs of the employers (Langenkamp, Costa, and Cheung, 2019). Similar to firms that do not employ AI, the information generated by predictive algorithms is likely to diminish the accuracy of information provided as it reflects the biased hiring choices made by employers rather than providing objective insights into the candidates' performance. This implies that

Hypothesis 3.2.1 *Resume screening through predictive algorithms (i) leaves gender equality in the applicant pool selected unaffected, compared to human recruiters, while (ii) leaving also the overall quality of the hiring process stable, compared to human recruiters.*

This hypothesis is in line with the existing research showing that predictive algorithms make biased hiring decisions because they learn to be biased from humans (Gonzalez et al., 2022; Black and van Esch, 2020; Gebru, 2020; Kochling and Wehner, 2020; Daugherty, Wilson, and Chowdhury, 2019; Silberg and Manyika, 2019; Bogen, 2019; O' Neil, 2016).

3.2.3 Firms use assessment software

Chapter 1 discussed assessment software as an AI tool used to assess job applicants based on their performance in cognitive tests, interviews, and chats (Li et al., 2021). Assessment software collects extensive information regarding the productivity of individual job applicants and utilizes this data to determine suitable candidates for hiring (Daugherty, Wilson, and Chowdhury, 2019). Consequently, assessment software does not rely on biased information about groups to make inferences about individual job applicants' productivity. Instead, it provides employers with accurate information regarding the productivity of individual job applicants. Additionally, since assessment software evaluates job applicants independently of past employees' data, it avoids perpetuating employers' biased beliefs and biased hiring decisions. Therefore, it is reasonable to argue that, unlike firms that do not employ AI or use predictive algorithms, the use of assessment software likely increases the accuracy of the information. As a result, employers rely on the accurate information provided by assessment software regarding candidates' performance when making hiring decisions, rather than biased information influenced by gender stereotypes. In other words, the use of assessment software in hiring allows firms to ensure meritocracy in their hiring process. Here, I treat - as from Castilla (2008) - meritocracy as "a process in which merit is somehow measured and then compensated. Meritocracy is thus one possible way of assigning rewards". Hence, it follows that

Hypothesis 3.2.2 *Resume screening through assessment software (i) increases gender equality in the applicant pool selected through meritocracy, compared to human recruiters and predictive algorithms, thereby (ii) improving the overall quality of the hiring process, compared to human recruiters and predictive algorithms.*

3.3 Setting and method

3.3.1 Setting

My data come from a private company in Italy. The company in 2022 was looking to hire a worker for an international sales and purchasing agent, which indicates the person responsible for the sales, purchasing and shipment of all products, materials and supplies internationally. The company never used assessment software or any other type of AI tool in hiring. The position requires at least a bachelor degree and experience with English and Spanish language. Further, the job advertisement states a preference for motivated workers, who are keen to improving their skills and knowledge on the job. Like other firms, my data provider faces challenges in identifying and hiring female applicants. In fact, as figure 3.1 shows, since 2014, female candidates comprise more than 60% of applications but only 33% of hires for the international sales and purchasing agent.

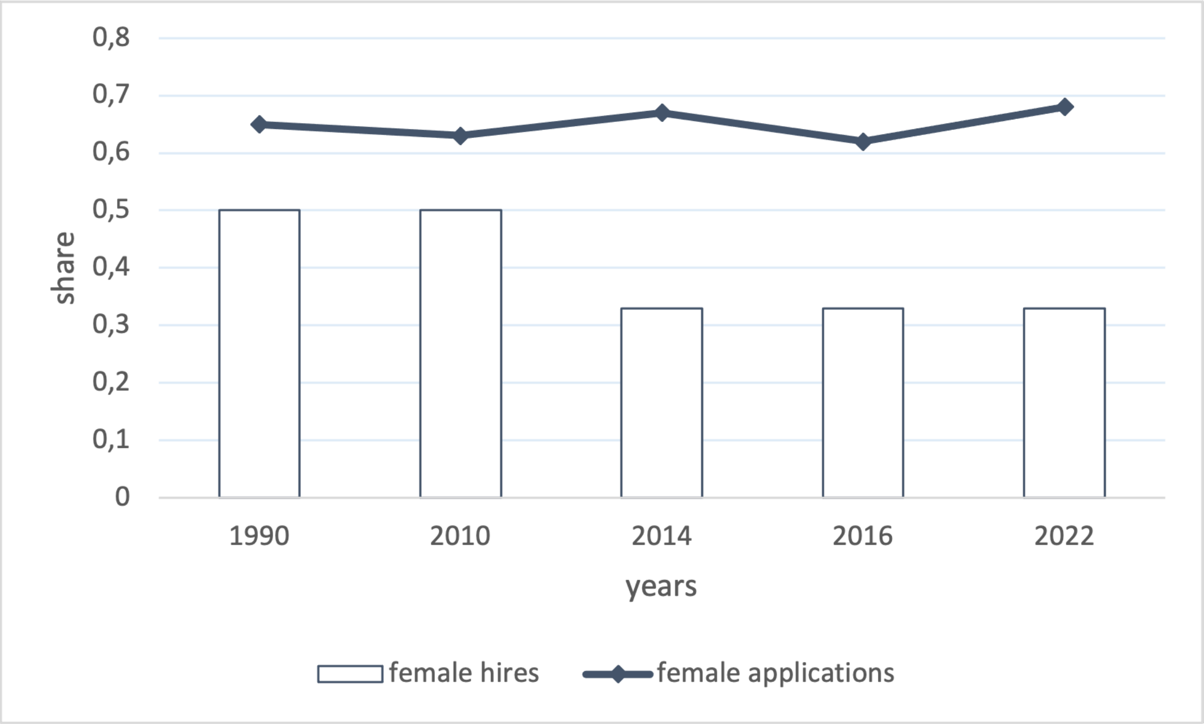


Figure 3.1: Firm's share of female hires and applications for the international sales and purchasing agent position over the years

The firm under study received approximately 200 applications for the international sales and purchasing agent position it was hiring for. Interview slots are very few and the time dedicated to resume screening is scarce because both are conducted by employees who are not HR professionals but administrative managers and accountants. They reduce the time devoted to their job in order to screen resumes and interview job candidates. Therefore, the firm under study is usually extremely selective in deciding whom to interview. In fact, it rejects on average 97% of the applicants prior to interviewing them. Further, this rejection decision is made only on the basis of resume screening.

Given the volume of job applications the firm receives and the little time recruiters can devote to resume screening, recruiters may easily rely on ascriptive caracteristics and heuristics to make their selection choices (Rivera, 2015). Therefore, they may easily end up discriminating against a certain group of job candidates (e.g., men or women). Further, recruiters may easily make mistakes and pass candidates who are not qualified enough for the job, while letting candidates who perfectly match firm's demand go.

Applicants qualifications and diversity

In this paper, first, I focus on whether with assessment software and predictive algorithms firms can improve gender equality in their selected applicant pool through meritocracy. Second, I focus on whether with assessment software and predictive algorithms firms can improve the quality of their selection decisions, measured by how the qualifications of job applicants selected for being interviewed matches firm's demand.

My firm faces a key challenge in identifying qualified workers because after usually rejecting 97% of the applicant pool as not qualified, yet about 50% of the new hires who are deemed qualified workers result in dismissals or resignations. Therefore, there is indeed scope for improving the selection process of this firm by extending selection opportunities to a more diverse set of applicants.

In summary, I investigate in the firm under study whether assessment software and predictive algorithms can (i) increase through meritocracy the diversity of the selected applicant pool, as defined by the share of female applicants shortlisted to be interviewed, thereby (ii) improving the overall quality of the hiring process as defined by the likelihood that candidate's qualifications match firm's demand.

3.3.2 Method

Data

I have data on 197 job applications from March 2022 to April 2022 for a international sales and purchasing agent position, as described in Table 3.1.

Table 3.1: Summary statistics

	Mean	SD
Female (=1)	0.68	0.47
Ethnic minority (=1)	0.22	0.42
B.A. or Master's Degree (=1)	0.73	0.44
Proficient in English (=1)	0.98	0.13
Proficient in Spanish (=1)	0.70	0.46
Keen to learn on the job (=1)	0.15	0.36
Observations	278	

NOTES: This table shows applicants' demographic characteristics, education histories, and proficiency in English and Spanish. Further, it shows the share of applicants who are keen to learn on the job. English and Spanish proficiency, and motivation to learn on the job were explicitly stated in the job advertisement as necessary requirements to be hired. All data come from the firm's application records. The number of observations refers to each item in job applicants' resumes.

The same application pool is screened by the assessment software and the predictive algorithm I developed, and by the recruiters.

To transform the raw information on job applicants' resumes into usable inputs for the assessment software and predictive algorithm, I create a series of dummy variables that serve as characteristics of each job applicant. Each variable is presented in table 3.1.

3.3.3 Assessment software and predictive algorithm

Here I describe how I construct the selection criteria based on a filtering tool, which I developed thanks to the advice of a major Italian AI company that creates hiring algorithms for firms. The tool I construct is based on the technology used by the most sold assessment software and predictive algorithm the company is commercializing in 2022.

Assessment software

I estimate the applicants shortlisted to be interviewed, conditional on possessing the requirements stated in the job advertisement as necessary to be hired. Applicants are selected according to the following strategy:

$$S_i = \begin{cases} 0, & \text{if } I(X_i) = 1 \\ 1, & \text{otherwise} \end{cases} \tag{3.1}$$

with S_i denoting whether the job applicant is shortlisted to be interviewed, and $I(X_i)$ indicator variable that indicates whether the job applicant possesses all job requirements. $I(X_i)$=1 if the job applicant is proficient in both English and Spanish, and has at least a B.A. Degree.

The model is designed to maximize qualifications, and have no additional preference related to diversity. Any gender difference in the applicant pool depends on job applicants' qualifications because I construct the algorithm without imposing any condition on gender.

While I eventually theoretically expect the algorithm to outperform human recruiters in both diversity of the selected applicant pool and qualifications of the selected applicants, it is unclear whether empirically there should be more or less diversity in the selected applicant pool. In particular, the impact of the algorithm on the selected applicant pool's diversity depends on the structure of the data, *i.e.*, on the characteristics of the applications that my firm receives.

Predictive algorithm

I predict the applicants shortlisted to be interviewed, conditional on being as similar as possible to whom was occupying the position of international sales and purchasing agent within the firm under study in the last 10 years. Applicants are selected according to the following strategy:

$$S_i = \begin{cases} 0, & \text{if } I(K_i) = 1 \\ 1, & \text{otherwise} \end{cases} \tag{3.2}$$

with S_i denoting whether the job applicant is shortlisted to be interviewed, and $I(K_i)$ indicator variable that indicates whether the job applicant has the same characteristics of the workers occupying the position in the past. $I(X_i)$=1 if the job applicant is proficient in both English and Spanish, has at least a B.A. Degree, and has been responsible for the marketing division during their career.

The model is designed to maximize qualifications, and have no additional preference related to diversity. Any gender difference in the applicant pool depends on job applicants' qualifications or other patterns the algorithm finds in the data because I construct the algorithm without imposing any condition on gender.

While I eventually theoretically expect the algorithm to outperform human recruiters in qualifications of the selected applicants, it is unclear whether empirically there should be more or less diversity in the selected applicant pool. In particular, the impact of the algorithm on the selected applicant pool's diversity depends on the structure of the data, *i.e.*, on the characteristics of the applications that my firm receives and on the characteristics of the workers that had occupied the position within the firm in the past 10 years.

3.3.4 Human recruiters

I evaluate whether my algorithms outperform human recruiters in both diversity of the applicant pool selected to be interviewed and qualifications of the selected job applicants. To do so, I compare qualifications and diversity in the three different applicant pools selected respectively by human recruiters, by the predictive algorithm and by the assessment software. To do so, I collect the list of resumes selected by human recruiters to be interviewed during their resume screening activity.

To compare how the algorithms and recruiters differently affect (i) diversity as defined by the share of female applicants shortlisted to be interviewed, and (ii) quality as defined by the likelihood that candidate's qualifications match firm's demand for the international sales and purchasing agent position, I estimate the following linear probability models

$$pr(female) = \alpha + \beta method + \Theta_i \gamma + \epsilon_i \tag{3.3}$$

$$pr(qualified) = \alpha + \beta method + \Theta_i \gamma + \mu_i \tag{3.4}$$

with Θ_i individual controls available in the resume (age and ethnic minority)[2], *method* dummies identifying whether the applicant was shortlisted by the predictive algorithm, the assessment software, or human recruiters[3]. The probability that the applicant is qualified is equal to 1 if the applicant meets the demand for qualifications needed to be hired by the firm, i.e., has at least a bachelor degree, a fluent knowledge of English and Spanish language, and is keen to learn on the job.

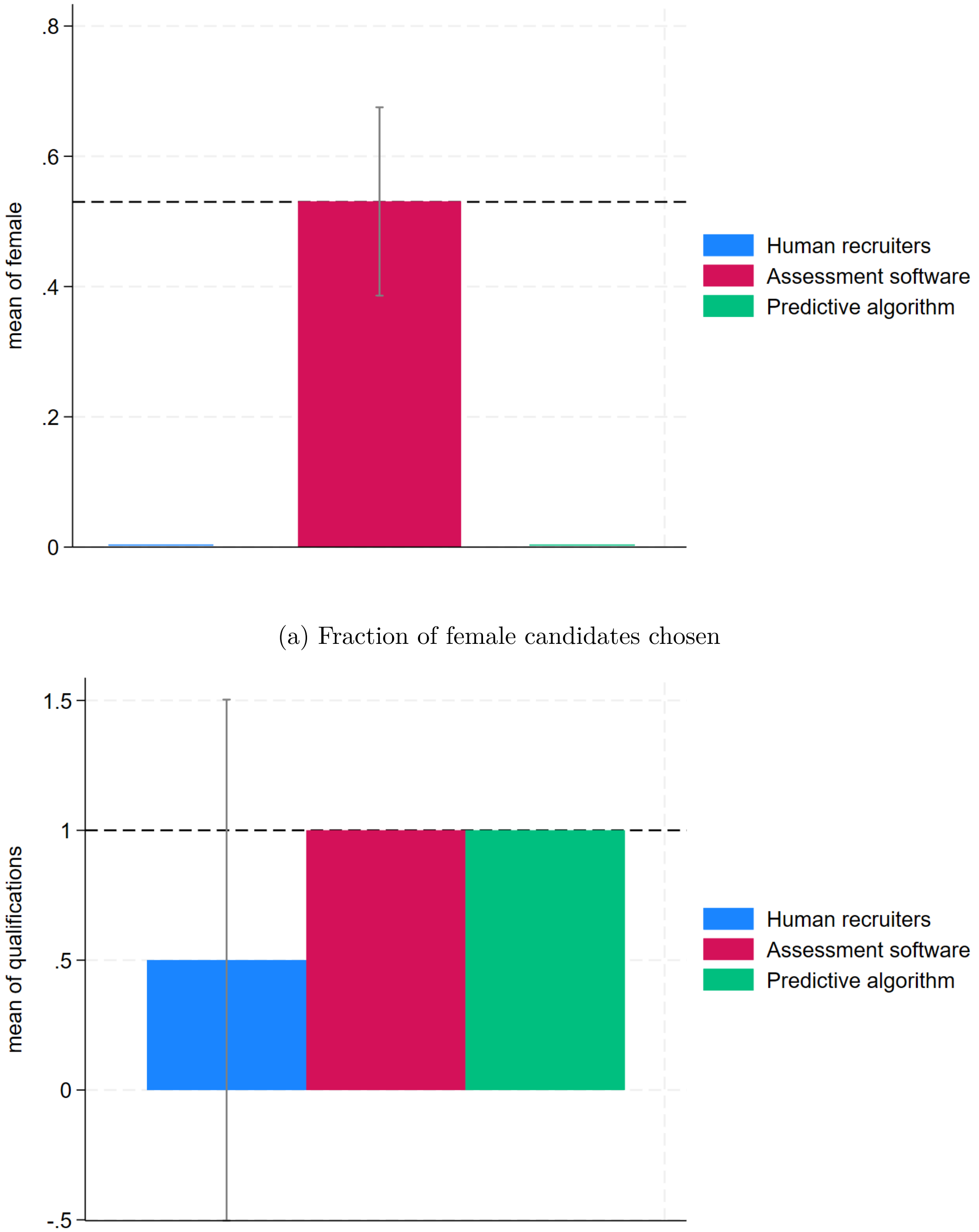
3.4 Results

3.4.1 Descriptives

Before presenting the results of the estimated linear probability model, figure 3.2 shows the fraction of female and qualified applicants shortlisted by each screening method (human recruiters, assessment software, and predictive algorithm).

[2] One may think to include the name of the university where the job applicant have studied as a control variable, since it signals high/low socio-economic status and high/low quality of education received. However, in a context such as Northern Italy, where my firm is located, it is unlikely that the university name matters because almost all universities in Northern Italy are public. The only university that would make the difference is Bocconi but there are no applicants in my data who have studied at Bocconi

[3] It is possible that the same applicants are selected with all the methods



(a) Fraction of female candidates chosen

(b) Fraction of qualified candidates chosen

Figure 3.2: Descriptive results

The top panel of figure 3.2 shows how often female candidates were chosen in each method. The results for the predictive algorithm show that the algorithm did not do significantly better than human recruiters (this follows from the observation that 0 female candidates were chosen by both methods). The assessment software, instead, does allow the firm to make more diverse selections (53% female candidates were chosen under this method).

The second panel of figure 3.2 shows how often qualified candidates were chosen in each method. The results for the assessment software and the predictive algorithm show that both algorithms did significantly better than human recruiters (this follows from the observation that 100% qualified candidates were chosen by both methods). Human recruiters, instead, performed as good as random (50% qualified candidates were chosen under this method).

3.4.2 Methods differences: test of the hypotheses

Table 3.2: Probability that a female applicant or a qualified applicant is selected for the interview by recruitment strategy

	Female applicant	Qualified applicant
Assessment software	0.79 (0.175)***	0.99 (.003)***
Predictive algorithm	0.22 (0.174)	1.01 (.010)***
N	278	278

***p<0.01, **p<0.05, *p<0.1

NOTES:Coefficient come from equations 3.3 and 3.4. Control variables include applicants' age and ethnicity for equation 3.3. Control variables include applicants' age gender and ethnicity for equation 3.4.

3.4.3 Diversity

I begin by assessing the different impact of the algorithms and recruiters on the diversity of the applicant pool selected to be interviewed. This is done through equation 3.3. This analysis is straightforward in the sense that I observe demographic covariates such as gender for all applicants, so that I can easily detect differences in the composition of applicants selected by the algorithms and by recruiters.

I begin by considering the gender composition of selected applicants. As presented in Table 3.1, at baseline 68% of applications in my data are female. Table 3.2 shows that relative to human recruiters, assessment software increases the probability of female applicants to be selected. In particular, the probability that a female applicant is selected increases by 79 percentage points. Also, table 3.2 shows that predictive algorithms do not differ from human recruiters in the probability of selecting female applicants (this follows from the observation that 0.22 falls within the 95%-confidence intervals for predictive algorithms).

3.4.4 Quality

I continue by assessing the different impact of the algorithms and recruiters on the qualifications of the applicant pool selected to be interviewed. This is done through equation 3.4. This analysis is also straightforward in the sense that I observe qualifications for all applicants, so that I can easily detect differences in how the selected job applicants meet the demand for qualifications needed to be hired by the firm. Table 3.2 shows that relative to human recruiters, both assessment software and predictive algorithms increase the probability of selecting qualified applicants to 1.

In sum, hypothesis 3.2.1 is partially confirmed by the analyses. Resume screening through predictive algorithms (i) leaves gender equality in the applicant pool selected unaffected, compared to human recruiters, while (ii) increasing the overall quality of the hiring process, compared to human recruiters. While hypothesis 3.2.2 is empirically confirmed. Resume screening through assessment software (i) increases gender equality in the applicant pool selected through meritocracy, compared to human recruiters and predictive algorithms, thereby (ii) improving the overall quality of the hiring process, compared to human recruiters. The probability to select qualified workers is the same as with predictive algorithms.

3.5 Discussion and conclusion

This chapter aims at contributing to the existing research on AI and how it affects gender inequality in hiring by shedding new light on how assessment software and predictive algorithms that automate the hiring process affect diversity and qualifications of the applicant pool selected. While a growing body of research has underlined the limitations of predictive algorithms in hiring and some potentiality of assessment software, this chapter takes a step further and highlights the different role of assessment software and predictive algorithms in affecting hiring outcomes under a comparative perspective. I show that by using assessment software to fully automate hiring decisions, not only do firms enhance gender equality in their hiring outcomes through meritocracy, but they also improve the overall quality of the hiring process, compared to the firm's current hiring practices.

A natural interpretation of this result is that human recruiters choose to minimize the risk of making more diverse hires at the expense of both quality and diversity. By showing that assessment software increases hiring quality but also diversity, I provide evidence that human recruiters do not make efficient hiring, ending up selecting weaker candidates over stronger candidates. Further, I show, in line with existing research, that although predictive algorithms improve the efficiency of the hiring process by selecting highly qualified job applicants, compared to human recruiters, they do not bring any improvement in the gender equality of firms' new hires.

My analysis, however, is limited by the fact that I only focused on the resume screening stage of the hiring process without exploring the use of assessment software and predictive algorithms during interviews and their impact on gender inequality in the final hiring decision. This limitation restricts the scope of my findings. Future studies should consider investigating this direction of research to gain further insights.

The most important concern raised by my analysis is that the results are based on the pattern of applicants that the algorithms happens to see in my data. If a different set of applicants had applied to my firm then it is possible that my results would change. A follow up to this study would aim at considering how algorithms and humans behave when the qualifications of applicants of different groups is changing over time.

Second, the possibility of omitted variables can lead to biases in my results. The algorithms are based on the characteristics of job applicants that I observe and may, therefore, overlook unobservable factors that human recruiters may consider when deciding whether a job applicant would make a successful hire. Such unobservable factors comprise the biases that human recruiters apply to the candidate evaluation process, such as previous experience or beliefs. To accurately assess the extent to which assessment software differently affects the qualifications level and diversity of the selected applicant pool compared to human recruiters, it would be ideal to accurately measure the level of bias that human recruiters include in their evaluation process. This would allow me to precisely quantify the magnitude of the effect that assessment software has in reducing bias in hiring. However, since I cannot accomplish this, my results may either overestimate or underestimate the potential of assessment software in improving gender bias in hiring when compared to human recruiters. Chapter 4 draws on such a limitation of the current chapter to develop a study about how assessment software and predictive algorithms can eventually improve the quality and equality of recruiters' hiring choices. In chapter 4, instead of asking whether full algorithmic hiring would lead to better outcomes, I ask whether firms can improve hiring quality by relying on algorithmic recommendations. The approach in chapter 4 is complementary to this chapter in that it does rely on modeling the human decision.

In spite of the limitations that constrain my results, my analyses show that when adopted by a firm, assessment software may identify highly qualified and gender diverse job applicants, who may be overlooked by human recruiters. Such changes brought about by assessment software may be incorporated into future predictive algorithms recommendations as they enter the training data. The magnitude of such a feedback loop across AI-types deserves future scrutiny.

4

AI and employers' choices (coauthored with Arthur Schram)

Abstract

Organizations increasingly use AI to assess job candidates, raising concerns about potential discrimination. This paper explores the consequences that the use of AI in candidate evaluation may have for discrimination and experimentally tests the theoretical predictions. Our model is based on the literature on information cascades. Two types of candidates are considered, one is randomly chosen to have a generic (but unknown to the employers) advantage in productivity. Each candidate also has a randomly assigned private productivity component. Employers sequentially choose between one of each candidate, receiving independent signals of candidates' productivity components and in some environments, observing previous employers' choices. We introduce AI's information on private productivity, which can be (i) none, (ii) an unbiased signal coming from assessment software, or (iii) a knowingly biased signal coming from predictive algorithms. Theoretical and experimental results show that (ii) and (iii) can break information cascades with positive probability, with AI improving choosers' decision-making even when providing biased information.

4.1 Introduction

In chapters 2 and 3, we learned that AI, particularly through assessment software is likely to reduce gender inequality in hiring, especially when it is granted full autonomy over hiring decisions. In order to complete the picture that this thesis aims to provide about how AI affects gender inequality in the labor market, one last step is missing. How does assessment software and predictive algorithms affect gender inequality in hiring when they are paired with human recruiters? Does the information that assessment software and predictive algorithms provide affect how recruiters make their hiring choices? Asking these questions is extremely relevant because how human recruiters receive, elaborate on, and make use of assessment software's and predictive algorithms' suggestions in hiring determines how strong the effect of AI can be in reducing or increasing gender inequality in the labor market. In particular, if recruiters deviate from the information provided by assessment software, while sticking to the information provided by predictive algorithms because it reflects human biased preferences, the effectiveness of assessment software in reducing gender inequality is undermined.

To investigate how assessment software and predictive algorithms interact with human biases in the labor market, the chapter is complementary to chapter 3 in that it completes the analysis of how assessment software and predictive algorithms affect quality and diversity of firms' hiring choices. Instead of assessing the impact of AI when it automates the hiring process, we model human decisions and study the impact of assessment software and predictive algorithms when they provide recruiters with hiring recommendations. For this purposes, we develop a theoretical model of employers' hiring choices when the assessment software or the predictive algorithm is available and subsequently test the model's predictions with an online experiment.

An important hypothesis we develop and test is that both assessment software and predictive algorithms can be an aid in avoiding that a labor market evolves into an information cascade. An information cascade refers to a setting where employers make hiring decisions based on historical decisions of previous employers, while ignoring private information they might have (Bikhchandani et al., 1992; Banerjee, 1992; Anderson and Holt, 1997). This could lead, for example, to a situation where employers prefer a man over a woman for a job because, historically, men have always done that job. In an information cascade, employers (rationally) choose the man even if they receive credible information that a female job candidate is better suited for the job than the male candidates being considered. The reason is that past choices signal that former employers had about the relative qualities of men and women. Though they may arise out of rational behavior, the aggregate outcome may be that such an information cascade occurs even though men are not actually better at a job than women.

The literature on labor markets and organizations has produced many insightful studies on how employers make hiring decisions when information cascades my appear. Anderson and Holt (1997) illustrate in the laboratory that “if several potential employers do not hire a worker because of a poor interview performance, a subsequent employer may not hire the worker even if the employer’s own assessment is favorable, because of the unfavorable signals inferred by previous employers’ rejections” (Anderson and Holt, 1997, p.847). In a similar vein, Kübler and Weizsäcker (2003) and Oberholzer-Gee (2008) show that this kind of herd behavior can explain why long spells of unemployment may reduce the chances of re-employment because it is a signal of previous employers’ evaluations of the candidate under consideration. This phenomenon can be extended from the history of a particular worker to a specific *type* of worker. If several potential employers prefer to hire a man for a specific type of job than a woman, a next employer may take this into account when choosing between a man and a woman for the same type of job.

We hope to contribute to the existing literature not only by providing the first theoretical and empirical demonstration that assessment software and predictive algorithms matter in affecting employers’ prior beliefs and hiring decisions, but also by revealing new mechanisms that may subtly increase or decrease biases in hiring outcomes.

In fact, we find that assessment software and predictive algorithms may break information cascades, especially, predictive algorithms, *i.e.,* when AI provides *biased* information. That is, when AI systematically favors one candidate over the other. The underlying intuition is that such information may be strong enough to help the employer to ‘break out’ of the cascade. We show that employers adjust their decisions according to the information provided by assessment software and predictive algorithms about the job candidates' productivity, increasing their probability to make a 'correct' choice. Therefore, the chapter shows that assessment software not only reduces gender inequality and improves the overall quality of the hiring process when it automates hiring but also when it is paired with human recruiters. The same discourse on quality applies also to predictive algorithms, which improve the overall productivity of the new hires not only when they automate the hiring process but also when they are paired with human recruiters.

4.2 Theory

4.2.1 Model setup

There are two types of job candidates, A and B. N employers sequentially choose between hiring either a candidate of type A or one of type B. Employer $j \in \{1, ..., N\}$ chooses between A_j and B_j and all employers are searching for a similar position. Denote the expected productivity of A_j and B_j by $E\{P_j^A\}$ and $E\{P_j^B\}$, respectively.

Employer j chooses the candidate $c_j \in \{A_j, B_j\}$ with the higher expected productivity, yielding

$$c_j = \begin{cases} A_j, & \text{if} \quad E\{P_j^A\} > E\{P_j^B\} \\ B_j, & \text{if} \quad E\{P_j^B\} > E\{P_j^A\} \end{cases} \tag{4.1}$$

We assume that j randomizes if she is indifferent between the two candidates. A worker's productivity consists of two components. First, there is a common component $\kappa^k \in [-1, 1]$, $k \in \{A, B\}$ that describes a common productivity for every candidate of type k. We define $\kappa \equiv \kappa^A - \kappa^B$ as the relative productivity advantage of workers of type A. If $\kappa >$ ($<$)0 then workers of type A (B) tend to be better at the type of job under consideration[1]. If $\kappa = 0$ then both types are equally suited a priori. We assume that κ

[1] For example, let the job being recruited for be that of a professional football player, and let type A (B) be players below (above) 35 years old. Typically (but not necessarily always), type A candidates will be better suited than type B.

follows cumulative distribution function $F(\cdot)$, which is symmetric around $\kappa = 0$.

The second component in a worker's productivity reflects idiosyncratic productivity differences, denoted by $\tau_j^k \in [-1, 1]$, $k \in \{A, B\}$. This covers specific talents of the candidate concerned[2]. We again normalize by considering the difference $\tau_j \equiv \tau_j^A - \tau_j^B$ as the relative talent advantage of the worker of type A under consideration by employer j. We assume that each τ_j follows cumulative distribution function $G(\cdot)$, which is symmetric around $\tau_j = 0$.

The relative expected productivity of worker A is then given by

$$E\{\Delta P_j\} \equiv E\{P_j^A\} - E\{P_j^B\} = E\{\kappa\} + E\{\tau_j\} \tag{4.2}$$

and employer j chooses to hire worker A (B) if

$$E\{\Delta P_j\} > (<) 0 \Leftrightarrow E\{\kappa\} > (<) - E\{\tau_j\} \tag{4.3}$$

Because $F(\cdot)$ and $G(\cdot)$ are both symmetric around 0, a priori $E\{\Delta P_j\} = 0$ and the employer will randomize. The employer will, however, collect information. We consider two types of information. Both types yield signals about the relative productivity to be expected from candidates A_j and B_j.

4.2.2 Signals

First, the employer receives a private signal k_j about the realized κ. This signal is determined by a random draw ω from a distribution $F'(\cdot)$ that is symmetric around κ and independent across j, with

$$k_j = \begin{cases} 'A' & \text{if} \quad \omega \geq 0 \\ 'B' & \text{if} \quad \omega < 0 \end{cases} \tag{4.4}$$

Because of the symmetry of $F'(\cdot)$ around κ, the probability of signal $'A'('B')$ is larger

[2] To continue with the football example, Tom Brady (American Football) and Gianluigi Buffon (Football in the rest of the world) used to be type B players with a very large τ^B.

(smaller) than 0.5 when $\kappa > (<)0$. Without further information about κ, the employer holds expectation $E\{\kappa\} < 0$ if $k_j =' B'$ and $E\{\kappa\} \geq 0$ if $k_j =' A'$[3]. In that situation, the employer will simply choose in accordance with the private signal that she receives (because a priori $E\{\tau_j\} = 0$).

Second, the employer may use assessment software or predictive algorithms to obtain a better estimate of $E\{P_j^A\}$ and $E\{P_j^B\}$. For example, assessment software may use data from candidates' resumes to make such an estimate, while predictive algorithms may use data from previous workers employed within the firm to make this estimate. We consider two types of data. First, AI observes the history, h_j, of all previous employers' choices on job applicants' resumes, that is, $h_j = \{c_1, ..., c_{j-1}\}$. A summary statistic of these previous choices is the difference in the number of times that the A and B type was chosen, denoted by

$$h_j^{da} \equiv \sum_{k=1}^{j-1} (1_{c\kappa=A} - 1_{c\kappa=B}) \tag{4.5}$$

where $1_x = 1$ if x is true and $1_x = 0$, otherwise. If $h_j^{dA} > 0$, then workers of type A were chosen more often by previous employers than workers of type B. h_j^{dA} is informative about the relative popularity of types A and B for previous employers and therefore about the private signals they had received. In this way, assessment software and predictive algorithms may help employers obtain a better estimate of κ than from their own signal k_j alone. We will discuss this further, below.

The second way in which assessment software and predictive algorithms might inform employers about $E\{P_j^A\}$ and $E\{P_j^B\}$ is by providing information about the idiosyncratic differences τ_j. Assessment software, in fact, analyzes candidates' resumes, or evaluates job candidates through web-based interviews, chats and cognitive games (Li et al., 2021). Predictive algorithms analyze candidates' resumes and predict whom to hire based on whether job candidates present the same skills and characteristics as incumbent successful workers (Rhea et al., 2022). We model this by assuming that AI provide employer j with

[3] After a signal $k_j =' B'$, the employer expects $E\{\kappa\}^B \equiv E_F\{\kappa | k_j =' B'\} = \int \kappa dF(\kappa | \omega < 0) < \int \kappa dF(\kappa) = 0$. Similarly, a signal $k_j =' A'$ yields expectation $E\{\kappa\}^A \equiv E_F\{\kappa | k_j =' A'\} = \int \kappa dF(\kappa | \omega > 0) > \int \kappa dF(\kappa) = 0$

a signal t_j that is a noisy representation of τ_j, that is,

$$t_j = \tau_j + \epsilon_j, \quad \text{with} \quad \epsilon \sim G' \tag{4.6}$$

Where G' has mean b and standard deviation σ. The mean b represents a bias that AI might have towards candidates of type $A(b > 0)$ or $B(b < 0)$. We call assessment software the unbiased AI, i.e., if $(b = 0)$. Note that the expected signal t_j is increasing in τ_j and b. If $\sigma > 0$ then the signal is noisy and biased and we call predictive algorithm the biased AI. Predictive algorithms provide a biased signal that systematically favors one candidate over the other because this is a formal way of modeling the fact that predictive algorithms reproduce existing human gender biases (as we know from existing research - see Gebru (2020) or O'neil (2017)). Conversely, as we saw in chapter 3, assessment software provides a signal that is unrelated to candidates' gender, thus it provides an unbiased signal that may favor one candidate over the other with equal probability across candidates types.

4.2.3 Employers' choices

Using equations 4.3 to 4.6, it follows that employer j will choose candidate $A_j(B_j)$ if

$$E\{\kappa|k_j, h^{dA}\} > (<) E\{-\tau_j|t_j\} \tag{4.7}$$

We start with the l.h.s. and describe the updating pattern of κ, assuming $\tau_j = 0$; that is, assessment software or predictive algorithms only provide information about previous employers' choices. This places the model in the realm of the literature on information cascades.

Employer j, thus, observes a private signal k_j about the realized κ and information h_j^{da} about the relative popularity of A- and B-type workers among previous employers. She uses these to update her beliefs $E\{\kappa\}$, and chooses candidate $A_j(B_j)$ if $E\{\kappa\} > (<)0$. It is well known that information cascades can arise in this environment (Anderson and Holt, 1997). To see this, assume that employer 1 receives signal $k_1 =' A'$, which occurs with positive probability $1 - F'\{0\}$, even if $\kappa < 0$. As explained above, employer 1 then believes that $E\{\kappa\} > 0$ and chooses candidate A_1. Subsequently, assessment software or predictive algorithms inform employer 2 that $h^{dA} = 1$ and employer 2 concludes that her predecessor 1 had received signal $'A'$. Now assume that 2 receives the same signal, $k_2 =' A'$. Based on the signal alone, 2 would select A_2; this choice is reinforced by the fact that 1 had received the same signal. Recall that 2 randomizes with equal probability between A_2 and B_2 if she receives the opposite signal than employer 1. Next, let employer 3 learn that $h^{dA} = 2$ and, thus, deduce that both 1 and 2 had received signal $'A'$ or that 1 received signal $'A'$, while 2 received signal $'B'$ and randomly chose A. If she were to receive signal $k_3 =' A'$ as well, she would select A_3. Even if she were to receive signal $k_3 =' B'$, she would rationally know that A is more likely to be the better option (c.f. Appendix). Given that distributions F and F' are fixed, Bayesian updating would still yield expectation $E\{\kappa\} > 0$ and she would select A_3 irrespective of her own signal. All subsequent employers realize that only the first two decisions matter and all will believe $E\{\kappa\} > 0$ and choose candidate A_j, even if $k_j =' B'$. In this environment, information cascades may occur.

More generally, an information cascade will arise at any point where $|h^{dA}| \geq 2$. Such a cascade may lead to repeated selection of the more or of the less suitable type of candidate for the job, even when there are no individual differences $(\tau_j = 0, \forall j)$[4]. The possibility of information cascades occurring in this environment is formalized in Proposition 1 in Appendix A.

This reasoning shows that when assessment software or predictive algorithms only inform about previous employers' choices, they can lead to information cascades. This occurs because they simply replace the public decision making assumed in the traditional information cascades literature by another source of the same information. This is, however, a very limited approach to what assessment software or predictive algorithms may do. We now consider the case where assessment software or predictive algorithms provides also a signal t about the individual difference between candidates A_j and B_j (cf. equation

[4] The occurrence of cascades in such a setup is caused by the deterministic decisions that are assumed (e.g., in eq. (4.7)). As shown in Goeree et al. (2007), allowing for stochastic decisions provides a 'breakaway' that allows decision makers to escape from cascades. We investigate below whether AI can also play this role.

4.6). We ask how this might affect the occurrence and stability of information cascades, and how this depends on AI's bias b.

First, consider the role of the error term ϵ_j in equation 4.6. As $\sigma \to \infty$, the signal t becomes completely random. As a consequence, it follows from equation 4.7 that t may overturn any decision based on $E\{\kappa|k, h^{dA}\}$. This provides a way out of any information cascade. As a consequence, previous decisions can no longer be directly attributed to private signals k_j. Even when σ is small, however, previous decisions may be attributed to realized τ's and, thus, information cascades can be avoided.

Now, consider the role of the AI's bias b. The effect of the bias on the likelihood of getting out of an information cascade depends on which cascade occurs and which type is favored by the bias. For example, if previous employers' decisions have induced a series of employers to choose the B candidate, then it is more likely that the cascade will end when predictive algorithms are used and are biased in favor of type A than when they favor type B.

4.3 Experimental procedures and design

4.3.1 Procedures

600 participants took part in the experiment, which was implemented online through the Prolific platform in November 2022. Experimental instructions are presented in Appendix B.

Participation takes approximately 15 minutes and average earnings were £4. Earnings in the experiment are denoted by 'tokens', which are exchanged for pounds at the end of the experiment at a rate of £0.02 per token. Participants act as 'employers' (called 'choosers' in the experiment) who sequentially have to choose between a (virtual) 'employee' of type A or B. Each chooser is randomly appointed one candidate of each type and has to choose between those two candidates. The payoffs related to the chosen candidate are described below.

We enrolled choosers in batches of five. The participants in a batch simultaneously read the instructions. Subsequently, they are randomly assigned to a position in the sequence of five choices. After all five have made their choice, a second batch starts, for which the decisions of the previous batch are used in the way described below. This continues for a total of five batches, thus creating a sequence of 25 choosers[5].

4.3.2 Design

To start, one of the types is randomly, with equal probability, chosen to be the 'bonus type'. The bonus type is assigned a bonus value equal to $\kappa = 40$[6]. This assignment of the bonus type takes place before the first chooser in a sequence makes her choice. The assigned bonus type subsequently holds for all 25 choosers in the sequence.

When it is a chooser's turn in the sequence, she is first provided with a private signal regarding the bonus type. Details are provided below. Next, each candidate is randomly and independently assigned a discrete 'candidate value' $\tau_i \in [-100, 100]$ tokens. To make a choice, depending on the treatment (as discussed below) the chooser is provided with partial information about the value of each of the two candidates. Subsequently, she chooses one of the two candidates to determine her payoffs. The payoffs are equal to the τ_i tokens of the chip chosen plus 40 tokens if the chip chosen is of the bonus type. Figure 4.1 summarizes the design.

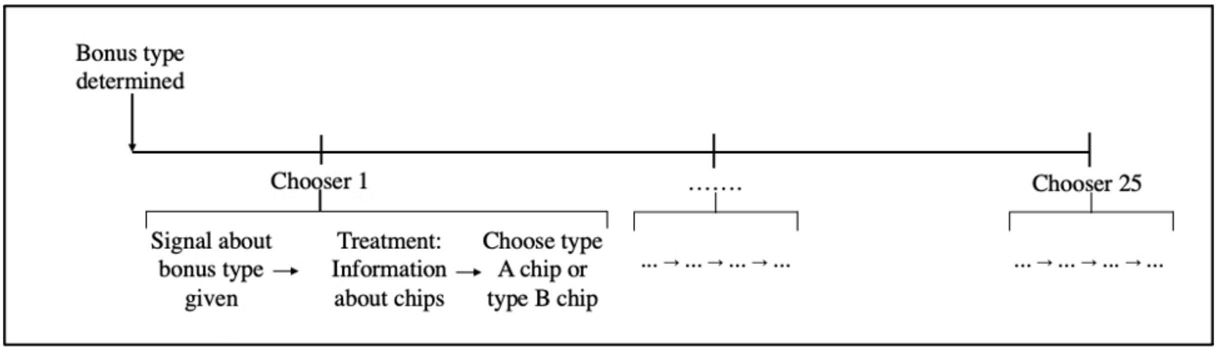

Bonus type determined
Chooser 1
....
Chooser 25
Signal about bonus type given →
Treatment: Information about chips →
Choose type A chip or type B chip
... → ... → ... → ...
... → ... → ... → ...

Figure 4.1: Timeline of the experiment
NOTES: The figure shows the timeline for one sequence of 25 choosers. Choosers sequentially pass though the same 3 steps. The treatment information varies across treatments but it is of the same type for all choosers in a sequence.

[5] This sequence length is comparable to what is used in experiments on information cascades. For example, sequences in Goeree et al. (2007) vary between 17 and 30 choices.

[6] Note that this reduces the distribution function F of our theoretical model to a 50-50 lottery between $\kappa = -40$ and $\kappa = 40$.

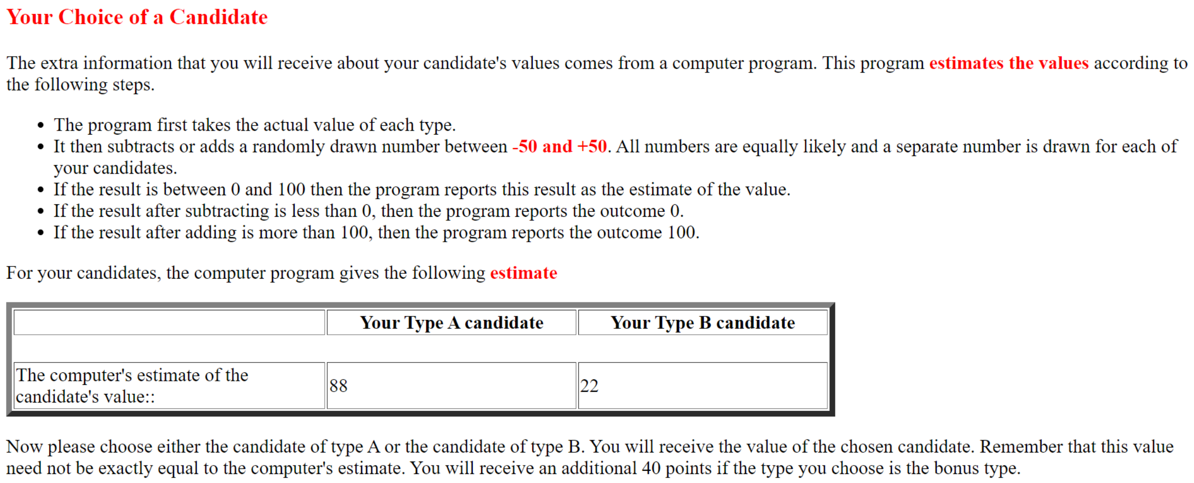
The private signal regarding the bonus types is presented in a way that is common practice in experiments on information cascades (e.g. Anderson and Holt (1997)). Each type is assigned a virtual vase with three balls. The vase of the bonus type contains two green balls and one blue ball. The vase of the other type has one green ball and two blue balls. On her monitor, the chooser sees both vases and the six balls, the colors of which are not shown. She may click on exactly one ball to reveal its color. Note that a revealed green (blue) ball implies that the vase from which it was selected in the bonus type with probability $\frac{2}{3}$ $\left(\frac{1}{3}\right)$. Finally, to simplify the treatment comparisons, we conducted the random draws of candidate values and AI signals once, before the first session, and separately for each position in the sequence. These same draws were used in all sequences of all treatments.

4.3.3 Treatments

Treatments involve the type of information given to the choosers. We vary this along two dimensions. The first concerns whether or not AI provides information about the type chosen by the choosers that preceded a position in the sequence. We either do not give any such information or we tell a chooser how often type A was previously chosen and how often type B was previously chosen (where chooser 1 is simply told that there were no previous choices). Recall that such 'historic' information may be informative about the bonus type because each previous chooser received an independent signal about this type.

In the second dimension, we provide the chooser with a signal concerning the values of the two candidates. This information can be of three kinds. In the first, there is no such information. The second type of information provides an unbiased signal about the two candidate values (assessment software's information), while the third provides a biased signal of the same (predictive algorithm's information). These signals are presented as follows. For each of the two candidate values, we randomly and independently draw a number from the discrete uniform distribution between -50 and +50. For the unbiased signal, we add the number drawn for the type-A candidate to the type-A candidate value and report the sum to the chooser. We do the same for the type-B candidate. For the biased signal, the distribution used for the type-A candidate is no longer uniform. Instead, the likelihood of a positive signal is twice as large as that of a negative signal. The distribution for the type-B candidate remains uniform. All of these procedures are common knowledge. Predictive algorithms provide a biased signal that systematically favors one candidate over the other because this is a formal way of modeling the fact that predictive algorithms reproduce existing human gender biases (as we know from existing research - see Gebru (2020) or O'neil (2017)). Conversely, as we saw in chapter 3, assessment software provides a signal that is unrelated to candidates' gender, thus it provides an unbiased signal that may favor one candidate over the other with equal probability across candidates types.

For the choosers in the experiment, the signals for the two candidate types are visualized as follows. The monitor first states that each type's actual value is used as a point of departure. Below this, we show a text stating that a random number between -50 and 50 is added. The text also explains what happens if the sum is smaller than 0 or greater than 100. Finally, a table shows the sum of the two numbers, for each type. Figure 4.2 shows the text and table choosers see on the monitor.



Your Choice of a Candidate

The extra information that you will receive about your candidate's values comes from a computer program. This program **estimates the values** according to the following steps.

- The program first takes the actual value of each type.
- It then subtracts or adds a randomly drawn number between **-50 and +50**. All numbers are equally likely and a separate number is drawn for each of your candidates.
- If the result is between 0 and 100 then the program reports this result as the estimate of the value.
- If the result after subtracting is less than 0, then the program reports the outcome 0.
- If the result after adding is more than 100, then the program reports the outcome 100.

For your candidates, the computer program gives the following **estimate**

	Your Type A candidate	Your Type B candidate
The computer's estimate of the candidate's value::	88	22

Now please choose either the candidate of type A or the candidate of type B. You will receive the value of the chosen candidate. Remember that this value need not be exactly equal to the computer's estimate. You will receive an additional 40 points if the type you choose is the bonus type.

Figure 4.2: Text choosers see on the monitor about how AI gives its signal

Table 4.1 summarizes our treatments. These treatments are varied between subjects, with four sequences of 25 choosers per treatments cell.

Table 4.1: Treatments

	History	
Information about candidate values	No	yes
No	Control	Information cascade
Unbiased	Unbiased without history	Unbiased with history
Biased	Biased without history	Biased with history

NOTES: The table summarizes our 3x2 treatment design. Treatments are varied between subjects, and we have data from four sequences (100 choosers) per treatment cell. 'History Information' involves the number of type A and type B choices by choosers that are earlier in the sequence.

Note that a comparison between treatments 'Control' and 'Information cascade' provide a replication test for the previous literature on information cascades (e.g., Anderson and Holt (1997), Goeree et al. (2007)). The only modification is that in our case information about previous choices is provided in aggregate form. The treatments involving information about the candidates reflect the distinct types of AI signals discussed in the theory section.

4.4 Hypotheses

Following from the theory section and the experimental procedures, AI may provide three types of information about the candidates that a chooser faces: (i) no information; (ii) an unbiased signal of the private productivity of the two candidates (assessment software's signal); (iii) a knowingly biased signal of their private productivity (predictive algorithm's signal). Our theoretical results show that both the assessment software's (unbiased) and the predictive algorithm's (biased) signals of candidates' private productivity break information cascades with positive probability. Our first hypothesis, thus, is:

Hypothesis 4.4.1 *Information cascades are less likely to occur when AI provides biased or unbiased information about candidates' private productivity.*

As a consequence, we expect a lower number of 'correct' decisions, choices of the more productive type, in the Information cascade treatment than in the control. Moreover, we expect biased and unbiased AI information to improve these decisions because they break some of the cascades. Our second hypothesis, thus, is:

Hypothesis 4.4.2.1 *In the Unbiased (assessment software) treatments, choosers probability to choose the more productive type is higher than in the Information cascade treatment.*

Hypothesis 4.4.2.2 *In the Biased (predictive algorithm) treatments, choosers probability to choose the more productive type is higher than in the Information cascade treatment.*

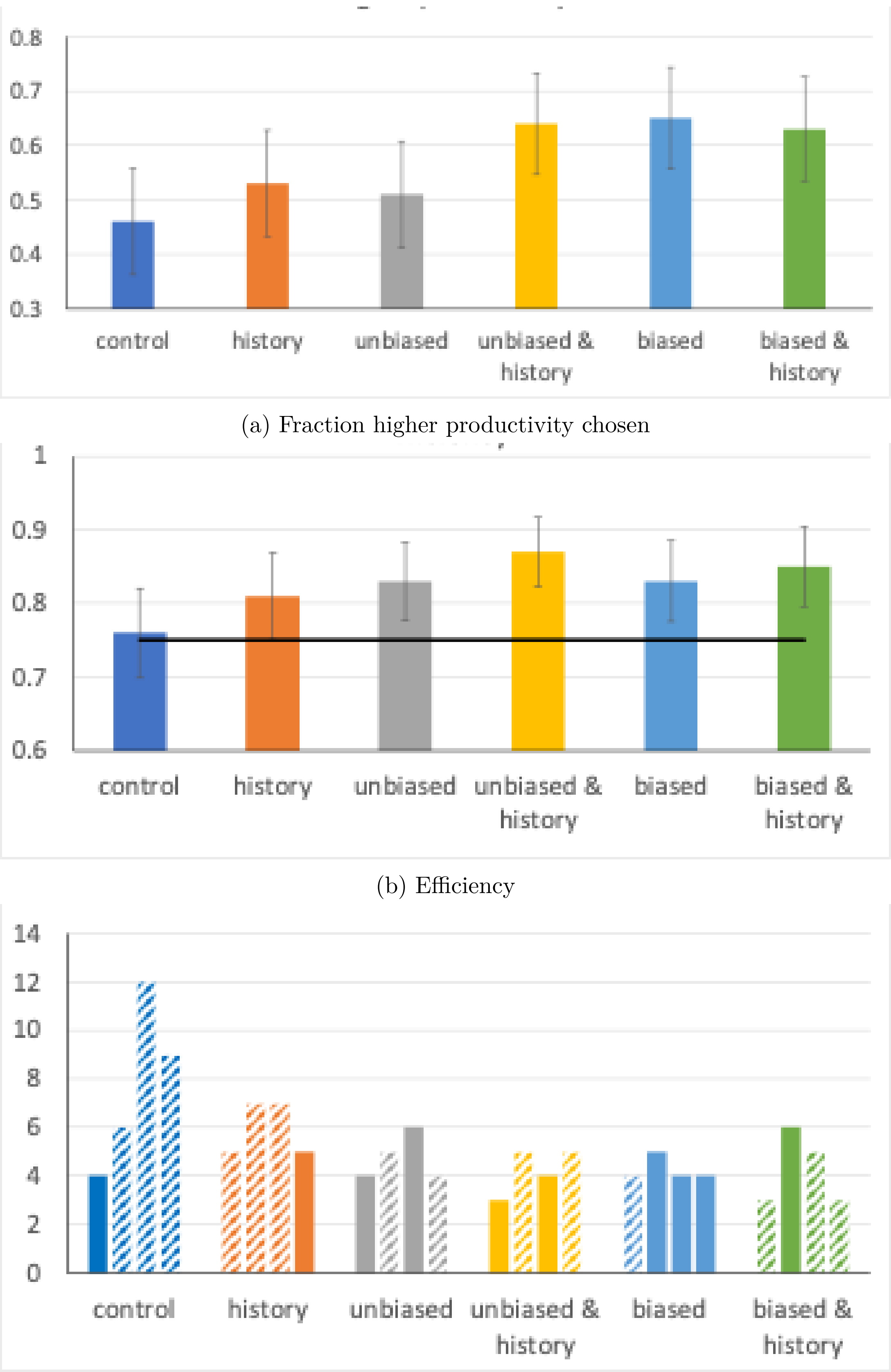
4.5 Results

4.5.1 Descriptive results

Figure 4.3 shows the basic results of our experiment. In particular, it shows per treatment the fraction of times that the candidate with higher productivity was chosen, the mean efficiency of the chosen candidates (defined as the ratio of the realized productivity and the higher of the two productivities), and the longest observed information cascade in each sequence.

The top panel of Figure 4.3 shows how often the option with the higher payoff (productivity) was chosen in each treatment. Note that a random choice between options would give a fraction of 0.5. The results for the control, the history treatment, and the unbiased AI treatment without history show that these participants did not do significantly better than they would have by choosing randomly (this follows from the observation that 0.5 falls within the 95%-confidence intervals for these treatments). The other three treatments (unbiased assessment software with history and biased predictive algorithm with and without history) do allow our participants to make better-informed decisions. We will discuss the consequences for our hypotheses below.

The middle panel of Figure 4.3 shows the mean efficiency of employers' choices. This is defined as the realized payoff divided by the higher of the two payoffs. Hence, the efficiency of an employer's choice is 1 whenever she chooses the better of the two options. Note that this occurs 50% of the time if an employer chooses randomly. With the value realizations that we drew, such a random choice would lead to an efficiency of approximately 0.75. This is indicated by a horizontal line in the figure. We observe that the participants



(a) Fraction higher productivity chosen

(b) Efficiency

(c) Longest information cascade

Figure 4.3: Descriptive results per treatment

NOTES: The top panel shows the fraction of cases where the candidate with the higher payoff was chosen. The middle panel shows the mean of the realized payoff divided by the higher of the two payoffs. The lower panel shows the longest information cascade observed in each of the sequences. Bars with a solid fill are cascades where the bonus type was chosen, bars with a pattern are cascades where the non-bonus type was chosen. Error bars show 95%-confidence intervals.

achieve a higher efficiency than the random-choice benchmark in all treatments except the control. Moreover, the best choices appear to be made when employers know the choices of their predecessors *and* assessment software provides an unbiased signal of the candidates' productivities.

Finally, the third panel of Figure 4.3 shows the length of the longest information cascade (that is, the longest sequence of participants choosing the same type) for each of the four sequences in every treatment. Here we find a remarkable and unexpected result. The two longest cascades are observed in the control treatment (12 and 9 employers in the third and fourth sequence, respectively). This result is unexpected because in this treatment, the only information that participants have is a signal about which type is the bonus type. Because this signal is independently drawn across participants, there cannot be any correlation in choices across participants. To illustrate what is going on, we consider the 12 participants (numbers 5-16 in the third sequence) that formed the longest 'cascade'. All of these participants chose the A-type candidate (while B was the bonus type). To obtain a signal, 9 (3) of these drew a ball from the A (B) urn. Now recall that a Bayesian updater will conclude that the chosen urn is (not) the bonus type if she draws a green (blue) ball. The first six participants in this sequence interpreted the signal is this way (that is, they chose candidate A after a green ball from the A urn or a blue ball from the B urn). Surprisingly, the next six employers did the opposite (choosing A after a blue ball from the A urn or a green ball from the B urn). We call the behavior by the first six of these 12 decisions 'Bayesian' and that by the others 'non-Bayesian'. The 'cascade' of 12 A choices in a row is then caused by (1) a coincidental sequence of six ball draws for Bayesians, followed by (2) a coincidental sequence of six non-Bayesians in a row. A closer look at our data for the control treatment reveals that it should not be very surprising that we observe six non-Bayesians. Of the 100 participants in this treatment, we classify no less than 73 as non-Bayesian[7]. This suggests that the signal that we provide about the bonus type is not used to update beliefs in the control, which explains why participants in the control treatment do not fare better than random choice would (Figure 4.3).

[7] We cannot classify participants in the other treatments as (non-)Bayesian because there the final choice of a Bayesian updater depends on more factors than just the urn chosen and the ball drawn.

Information cascades in the other treatments are shorter than in the control. Moreover, they appear to be shorter when more information is given than just the previous decisions by other employers (compare the history + (un)biased information treatments to the history treatment). We discuss the statistical significance of these differences below. Finally, recall that information cascades may select the good type (that is, the type with the bonus) or the bad type. The lower panel shows that in 14 of the 24 sequences that we collected, the longest cascade was of the 'wrong' type, in the sense that the bonus type was not selected. The equivalence of this phenomenon outside the laboratory would be a labor market where men are repeatedly hired even though women are on average more productive in the job concerned.

4.5.2 Treatment differences: testing our hypotheses

We use regression analysis to discern treatment differences and to formally test our hypotheses. First, consider hypothesis 4.4.1, that information cascades are more likely to be broken when there is biased (predictive algorithm) or unbiased (assessment software) AI information about the candidates. Recall from our theoretical model that, without further AI information, an information cascade is predicted whenever $|h_j^{dA}| \geq 2$. To test this, we created dummy variables that indicate whether the information about previous employers' choices revealed two more A choices than B choices or vice versa. We then apply probit regressions of the likelihood that an employer chooses type A on these two variables. We use data only from the treatments with information about previous employers' choices and disregard choices by the first two employers in a sequence (because no cascades is theoretically predicted before the third employer)[8]. A test of hypothesis 4.4.1 requires running this regression for the information cascade treatment –where a significant effect of the cascade dummies is predicted– and for the combined data from the two sessions with additional (biased and unbiased) AI information –where a smaller effect of the cascade dummies is predicted. Table 4.2 shows the results.

[8] For our regression analyses, we disregard 7 participants without a high-school diploma. Their choices perfectly correlate with choosing the A type.

Table 4.2: AI and information cascades

	Information cascades	History and additional AI information
# previous A choices - # previous B choices > 2	0.19 (0.08)**	-0.08 (0.07)
# previous A choices - # previous B choices < -2	0.27 (0.18)	0.01 (0.17)
N	90	182

NOTES: Cells show the estimated marginal effects of the historic information on the likelihood of choosing candidate A. Standard errors are in parentheses. '**' indicates statistical significance at the 5%-level.

Table 4.2 provides partial support for hypothesis 4.4.1. In particular, a history revealing that A was chosen at least twice more often than B by previous employers significantly increases the likelihood that an employer will choose A by 19 percentage points. When AI also provides other information, such information has no significant effect on the likelihood of choosing A. Combined, this provides support for the hypothesis. When employers are informed that the B type was previously chosen more often, however, this does not have the predicted negative effect on choosing A in the history only treatment. In fact, the marginal effect is positive, albeit insignificantly so. Adding other information does not change this null effect.

We now turn to hypotheses 4.4.2.1 and 4.4.2.2. These predict treatment differences in the extent to which employers hire the more productive employer (cf. Figure 4.3). To test these, we run a probit regression of hiring the more productive of the two candidates on treatment dummies and a set of employer background variables. Because the hypotheses do not distinguish between unbiased assessment software information with and without history, we pool these two treatments. The same holds for the two treatments with biased predictive algorithm information.

Table 4.3: Treatment and choosing the better candidate

	I	II
Information cascade treatment	0.08 (0.04)**	0.08 (0.04)**
Unbiased information treatments	0.11 (0.05)**	0.11 (0.05)**
Biased information treatments	0.18 (0.05)***	0.17(0.05)***
Age: 31-45		0.02(0.05)
Age: >45		-0.06(0.05)
Female		0.02(0.05)
Education: high school		0.00(0.04)
Education: master degree		0.08(0.07)
Education PhD or higher		0.17(0.14)
Employment: part time		0.01(0.05)
Employment: retired		0.12(0.11)
Employment: seeking opportunities		0.06(0.06)
Employment: prefer not to say		-0.02(0.07)
Works in human resources		0.02(0.05)
N	593	593
test Unbiased information = Information cascade	$\chi^2 = 0.53, p = 0.47$	$\chi^2 = 0.62, p = 0.43$
test Biased information = Information cascade	$\chi^2 = 9.72, p < 0.00 * **$	$\chi^2 = 7.85, p < 0.00 * **$

NOTES: Cells show the estimated marginal effects of the treatment dummies and demographic variables on the likelihood of choosing the more productive candidate. Standard errors are in parentheses. The benchmark categories absorbed in the constant term are the control treatment, age<31, males or prefer not to say, education with a bachelor degree, and employed full time. The tests in the last two rows are discussed in the main text. **/*** indicates statistical significance at the 5%/1%-level.

First note that it does not matter for the treatment effect whether we add demographic variables.

Hypotheses 4.4.2.1 and 4.4.2.2 predict that employers in the unbiased and biased treatments, respectively, are more likely to choose the better candidate than employers in the information cascade treatment. We test this with the χ^2 tests in the last two rows of Table 4.3. These test the following sets of hypotheses: $H_0^a : \beta_{unbiased} = \beta_{informationcascade}$ versus $H_1^a : \beta_{unbiased} > \beta_{informationcascade}$ and $H_0^b : \beta_{biased} = \beta_{informationcascade}$ versus $H_1^b : \beta_{biased} > \beta_{informationcascade}$. The results show that employers make better choices with biased (predictive algorithm) advice (confirming hypothesis 4.4.2.2) but not with unbiased (assessment software) advice (not supporting hypothesis 4.4.2.1). The result for biased (predictive algorithm) advice shows that the information provided by predictive algorithms is salient when choosers receive not only a private signal about job applicants' productivity but also information on how previous choosers evaluated job candidates.

4.6 Conclusion

In this chapter, we asked whether and how assessment software and predictive algorithms affect employers' decisions and human biases in the labor market, when they complement human recruiters in hiring. For this purpose, we developed a theoretical model of employers' hiring choices when such assessment software and predictive algorithms are available and subsequently tested the model's predictions with an online experiment.

Our theoretical model predicts that providing additional information with assessment software and predictive algorithms about the candidates will help to break cascades and improve decision making, even if this information is biased.

Our experimental results provide some support for the pattern formalized in hypotheses 4.4.2.1 and 4.4.2.2. We find that assessment software and predictive algorithms break information cascades with positive probability, particularly when the information provided is biased towards one of the candidates, thus coming from predictive algorithms. We find that predictive algorithms increase the probability that choosers make the correct choice (in the biased case). Though they make better choices with assessment software unbiased candidate information than without any such information, this difference is not statistically significant (Table 4.3). We cannot say at this point whether this may be attributed to the specific parameters we chose in our experiment (note, however, that our theoretical predictions do hold for these parameters). It is noteworthy that predictive algorithms biased information can lead to more extreme differences between candidates, and therefore may be more likely to break a cascade.

A more complete overview of what causes information cascades to appear and what makes them break requires more research. Our study aims at providing some basic insights into likely determinants. Two conclusions that stand out are (1) information cascades may improve decision making if employers do not carefully consider their private signals; (2) hiring decisions may be even further improved if AI information about specific candidates is used, because these help to break cascades, even (and perhaps especially) if such information is biased. We believe that the latter conclusion adds a novel element to the discussion about the influence of assessment software and predictive algorithms in hiring.

Though there are obvious disadvantages to receiving biased advice, the observation that such advice can serve as a mechanism that allows one to break out of an information cascade points to a possible advantage.

Indeed, this chapter suggests important and controversial implications of pairing assessment software and predictive algorithms with human recruiters in hiring that need further discussion, especially when compared and contrasted with the results presented in Chapters 2 and 3. Chapter 5 aims to conclude by engaging in such a discussion and elaborating on what we have learned from the three studies.

Appendix A

We start with establishing that providing historic information leads to information cascades. Define $q = P(k_j = A|\kappa > 0) = P(k_j = B|\kappa < 0) > 0.5$. q is the probability that the ‘correct’ signal is given. The probability that the ‘wrong’ signal is given (signal ‘A’ if $\kappa < 0$ or ‘B’ if $\kappa > 0$) is then $1 - q$. We then have

Proposition 4.6.1 *If AI provides (only) information h^{dA}, then the probability of an information cascade converges to 1 as the number of employers increases. The probability of a cascade where the incorrect candidate type is chosen is strictly positive for $q \in (0.5, 1)$.*

Proof 4.6.1 *Let $p_j \equiv P(\kappa > 0|h^{dA})$ denote the common knowledge posterior probability that $\kappa > 0$ (and thus A is the preferred candidate) given AI's signal h^{dA}, with $p_1 = 0.5$. Given p_j and private signal $k_j = A$, employer j believes that $\kappa > 0$ (and therefore that A is the preferred candidate) with probability $\pi_j^A(p_j)$, given by*

$$\pi_j^A(p_j) \equiv P(\kappa > 0|k_j = A, h^{dA})$$

$$= \frac{P(\kappa > 0)P(k_j = A, h^{dA}|\kappa > 0)}{P(\kappa > 0)P(k_j = A, h^{dA}|\kappa > 0) + P(\kappa \leq 0)P(k_j = A, h^{dA}|\kappa \leq 0)}$$

$$= \frac{P(k_j = A|\kappa > 0) \times P(h^{dA}|\kappa > 0)}{P(k_j = A|\kappa > 0) \times P(h^{dA}|\kappa > 0) + P(k_j = A|\kappa \leq 0) \times P(h^{dA}|\kappa \leq 0)}$$

$$= \frac{P(k_j = A|\kappa > 0) \times \frac{P(\kappa>0|h^{dA})\times P(h^{dA})}{P(\kappa>0)}}{P(k_j = A|\kappa > 0) \times \frac{P(\kappa>0|h^{dA})\times P(h^{dA})}{P(\kappa>0)+P(k_j=A|\kappa\leq 0)\times \frac{P(\kappa\leq 0|h^{dA})\times P(h^{dA})}{P(\kappa\leq 0)}}}$$

$$= \frac{q \times p_j}{q \times p_j + (1-q) \times (1-p_j)} \tag{4.8}$$

where we use $P(\kappa > 0) = P(\kappa \leq 0)$ (because $\kappa = 0$ has zero measure) and the independence of signals k_j and h^{dA}. Likewise, given p_j and private signal $k_j = B$, employer j believes that $\kappa > 0$ (and therefore that A is the preferred candidate) with probability

$$\pi_j^B(p_j) \equiv P(\kappa > 0|k_j = B, h^{dA})$$

$$= \frac{P(\kappa > 0)P(k_j = B, h^{dA}|\kappa > 0)}{P(\kappa > 0P(k_j = B, h^{dA}|\kappa > 0) + P(\kappa < 0)P(k_j = B, h^{dA}|\kappa < 0)}$$

$$= \frac{P(k_j = B|\kappa > 0) \times P(h^{dA}|\kappa > 0)}{P(k_j = B|\kappa > 0) \times P(h^{dA}|\kappa > 0) + P(k_j = B|\kappa < 0) \times P(h^{dA}|\kappa < 0)}$$

$$= \frac{P(k_j = B|\kappa > 0) \times \frac{P(\kappa>0|h^{dA}) \times P(h^{dA})}{P(\kappa>0)}}{P(k_j = B|\kappa > 0) \times \frac{P(\kappa>0|h^{dA}) \times P(h^{dA})}{P(\kappa>0)} + P(k_j = B|\kappa < 0) \times \frac{P(\kappa<0|h^{dA}) \times P(h^{dA})}{P(\kappa<0)}}$$

$$= \frac{(1-q) \times p_j}{(1-q) \times p_j + q \times (1-p_j)} \tag{4.9}$$

It follows from $q > 0.5$ that $\frac{q}{1-q} > 1$. A simple computation then shows that $\pi_j^A(p_j) > p_j > \pi_j^B(p_j), \forall p_j$. In other words, any given common posterior probability following the AI signal is adjusted downward after signal $k_j = A$ and upward after signal $k_j = B$. See Goeree et al. (2007) for a similar line of argument for traditional information cascade models.

To derive the Nash equilibrium for this game, consider first employer 1. Note that $p_1 = 0.5$. It follows from equation 4.8 that

$$\pi_1^A(p_1) = \frac{0.5q}{0.5q + 0.5(1-q)} = q > 0.5 \tag{4.10}$$

and from equation 4.9 that

$$\pi_1^B(p_1) = \frac{0.5(1-q)}{0.5(1-q) + 0.5q} = 1 - q < 0.5 \tag{4.11}$$

In other words, after signal A the first employer will choose candidate A_1 and after signal B she will choose B_1. Now consider employer 2. If employer 1 chose candidate A_1, then $h^{dA} = 1$. This gives

$$p_2 = P(\kappa > 0|h^{dA} = 1) = \frac{P(h^{dA} = 1|\kappa > 0) \times P(\kappa > 0)}{P(h^{dA} = 1)} = \frac{0.5q}{0.5} = q \tag{4.12}$$

which gives (from equation 4.8)

$$\pi_2^A(p_2) = \frac{q^2}{q^2 + (1-q)^2} > \frac{q^2}{2q^2} = 0.5 \tag{4.13}$$

where the inequality holds because $q > 0.5$. Equation 4.9 gives

$$\pi_2^B(p_2) = \frac{(1-q)q}{(1-q)q + q(1-q)} = 0.5 \tag{4.14}$$

In other words, if employer 1 chooses A then employer 2 will choose A after signal A and is indifferent between the two candidates after signal B. We assume that an employer who is indifferent randomizes with equal probabilities between the two options.

Now, consider employer 3. Employer 3 observes $h^{dA} = 2$ if:

- *employers 1 and 2 both received signal A. If $\kappa > 0$ this occurs with probability q^2. If $\kappa < 0$ this occurs with probability $(1-q)^2$.*

- *employer 1 received signal A and employer 2 received signal B and randomly selected candidate A_2. If $\kappa > 0$ this occurs with probability $0.5 \times q(1-q)$. If $\kappa < 0$ this occurs with probability $0.5 \times (1-q) \times q$.*

This gives

$$\begin{aligned} p_3 &= P(\kappa > 0 | h^{dA} = 2) \\ &= \frac{P(h^{dA} = 2|\kappa > 0) \times P(\kappa > 0)}{P(h^{dA} = 2|\kappa > 0) \times P(\kappa > 0) + P(h^{dA} = 2|\kappa \leq 0) \times P(\kappa \leq 0)} \\ &= \frac{[q^2 + 0.5 \times q \times (1-q)]}{[q^2 + 0.5 \times q \times (1-q)] + [(1-q)^2 + 0.5 \times (1-q) \times q]} \\ &= \frac{0.5q(1+q)}{q^2 + (1-q)^2 + q(1-q)} \end{aligned} \tag{4.15}$$

Straightforward calculations show that $p_3 > q = p_2$. In other words, an employer being informed that the previous two employers chose the candidate of type A has a stronger prior in favor of A than an employer who is informed that a single previous employer chose type A. As a consequence, employer 3 will choose A_3 after signal A. After signal B, her updated probability of A being the better candidate is

$$\begin{aligned} \pi_3^B(p_3) &= \frac{(1-q)p_3}{(1-q)p_3 + q(1-p_3)} \\ &= \frac{\frac{1}{D}0.5q(1-q^2)}{\frac{1}{D}[0.5q(1-q^2) + q(D - 0.5q(1+q))]} \end{aligned}$$

$$= \frac{0.5(1-q^2)}{0.5(1-q^2)+q^2+(1-q)^2+q(1-q)-0.5q(1+q)}$$

$$= \frac{0.5(1-q^2)}{1.5(1-q)} = \frac{1}{3}(1+q) > 0.5, \forall q \in (0.5, 1] \tag{4.16}$$

where $D = q^2 + (1-q)^2 + q(1-q)$*. It follows that an employer 3 observing* $h^{dA} = 2$ *will choose candidate A even after receiving signal B. By extension, this holds for all subsequent employers.*

Now consider an employer $j1 > 3$ *who observes that equal numbers of previous employers chose A and B, that is,* $h^{dA}_{j1} = 0$*. Because of symmetry, it must hold that* $p_{j1} = 0.5$*. Employer j therefore faces the same decision as employer 1.*

Next, consider an employer j2 who observes $h^{dA}_{j2} = 1$*, with n+1 previous employers having chosen A and n having chosen B, while* $|h^{dA}_i| < 2, i = 1, ..., j2 - 1$*; that is, no previous employer faced a difference of more than 1 previous choices. This means that employer j2-1 must have faced equal numbers of previous A and B choices and chose A, that is, j2-1 was in the position of j1, putting employer j2 in the same position as employer 2.*

Finally, consider an employer j3 who is the first to observe $h^{dA}_{j2} = 2$*, with n+2 previous employers having chosen A and n having chosen B, while* $|h^{dA}_i| < 2, i = 1, ..., j3 - 1$*. This means that employer j3-1 must have faced* $h^{dA}_{j3-1} = 1$ *and chose A. That is, j3-1 was in the position of j2, putting employer j3 in the same position as employer 3. Therefore, j3 always chooses A and* $h^{dA}_i > 2, \forall i > j3$*.*

Ergo, as soon as $h^{dA}_i = 2$ *for some i, all subsequent employers choose A. Note that this justifies our focus on j2's whose predecessors had never observed* $h^{dA}_i = 2$*. Once* $h^{dA}_i = 2$ *has been observed, an information cascade in A occurs. The same holds for a cascade in B once* $h^{dA}_i = -2$ *has been observed. It is easy to see that the probability of not observing* $h^{dA}_i = 2$ *or* $h^{dA}_i = -2$ *converges to 0 as* $i \to \infty$*. In other words, in the limit employers end up in an information cascade.*

Such an information cascade may be 'correct' (A if $\kappa > 0$ *or B if* $\kappa < 0$*) or incorrect (otherwise). As shown by Goeree et al. (2007) in a similar context, the likelihood ratio of a correct versus incorrect cascade is* $\frac{q}{(1-q)}$*, which increases from* $\frac{1}{2}$ *(both cascades are* *equally likely) to 1 (incorrect cascades are impossible) as q increases from 0.5 to 1.*

□

Next, we consider the case where each employer receives both a private signal about the common value κ and a private signal (from AI) about the difference between her two candidates, τ. In addition, AI tells her the history of choices by previous employers in her sequence, h_i^{dA}. Propositions 4.6.2 and 4.6.3 establish that the probability of an information cascade converges to zero in this environment when the τ-signal is, respectively, unbiased (coming from assessment software) and biased (coming from predictive algorithms).

Proposition 4.6.2 *If AI sends an unbiased signal about candidates A_j and B_j, then the probability of an information cascade converges to 0 as the number of employers increases. A priori, A_j and B_j are equally likely to be chosen.*

Proposition 4.6.3 *If AI sends a biased signal about candidates A_j and B_j, then the probability of an information cascade converges to 0 as the number of employers increases. A priori, A_j and B_j are equally likely to be chosen.*

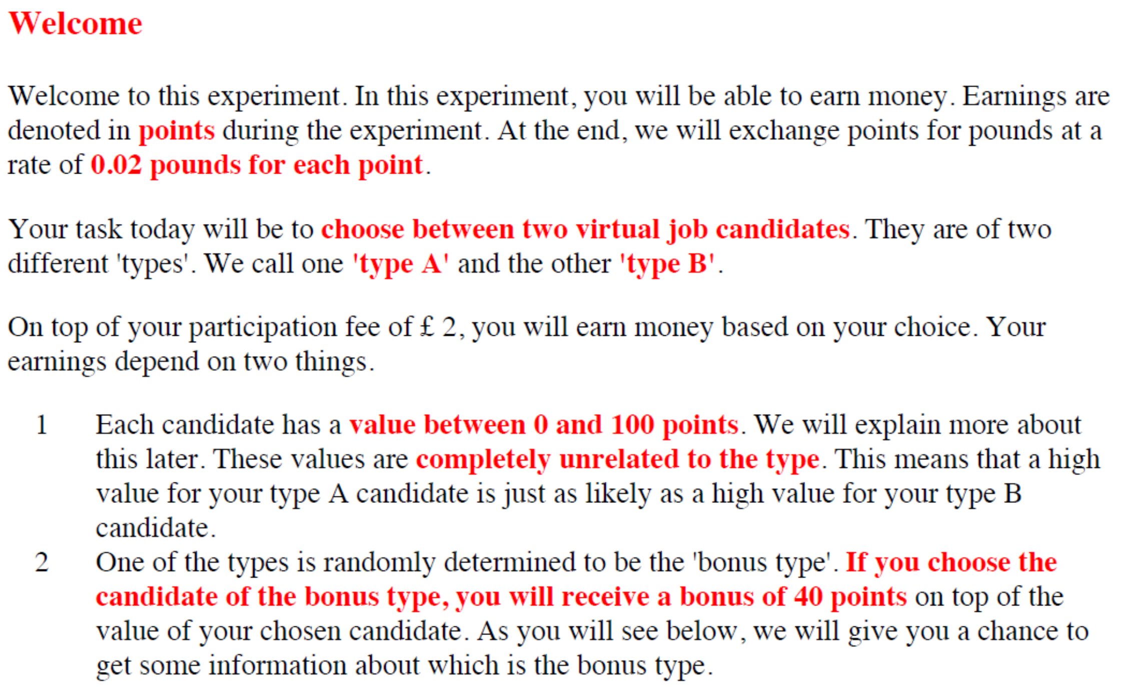
Proof 4.6.2 *The proof for both propositions is the same. Recall from equation 4.7 in the main text that an employer will choose A (B) whenever*

$$E\{\kappa|k_j, h^{dA}\} > (<)E\{-\tau|t_j\} \tag{4.17}$$

Note from the proof of proposition 4.6.1 that the l.h.s. of equation 4.17 A3 is strictly above –1 and below 1. Because the distribution G of signal t_j is unbounded, there is a positive probability that the r.h.s. exceeds in absolute value the l.h.s. This is enough to overturn an information cascade that may have arisen when new signals k_j cannot change the sign of $E\{\kappa\}$.

□

Appendix B



Welcome

Welcome to this experiment. In this experiment, you will be able to earn money. Earnings are denoted in **points** during the experiment. At the end, we will exchange points for pounds at a rate of **0.02 pounds for each point**.

Your task today will be to **choose between two virtual job candidates**. They are of two different 'types'. We call one **'type A'** and the other **'type B'**.

On top of your participation fee of £ 2, you will earn money based on your choice. Your earnings depend on two things.

1. Each candidate has a **value between 0 and 100 points**. We will explain more about this later. These values are **completely unrelated to the type**. This means that a high value for your type A candidate is just as likely as a high value for your type B candidate.
2. One of the types is randomly determined to be the 'bonus type'. **If you choose the candidate of the bonus type, you will receive a bonus of 40 points** on top of the value of your chosen candidate. As you will see below, we will give you a chance to get some information about which is the bonus type.

Summary

Thus, on top of your participation fee of £ 2, your earnings in tokens are as follows.

- If you choose your type A candidate and type A is the bonus type, you will earn 40 plus the value of your type A candidate.
- If you choose your type A candidate and type B is the bonus type, you will earn the value of your type A candidate.
- If you choose your type B candidate and type A is the bonus type, you will earn the value of your type B candidate.
- If you choose your type B candidate and type B is the bonus type, you will earn 40 plus the value of your type B candidate.

In what follows, we will explain each step of the experiment in more detail.

Sequences

You are grouped with 24 other participants in a **group of 25**. Each member of the group has to choose between one candidate of type A and one candidate of type B, but the candidates (and their values) are **different** from one participant to another. The **bonus type, however, is the same** for all 25 members of the group.

The members of the group will take turns in choosing between their two candidates.

First, participant 1 will decide, then participant 2, etc., up to participant 25. You are participant # 1 in your group.

Bonus Type

First, we will give you some information that may help you guess whether type A or type B is the bonus type that gives you 40 extra tokens.

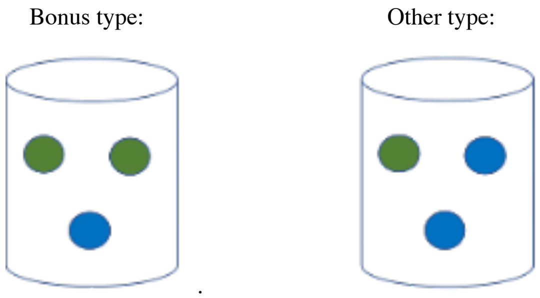
To do so, we create two virtual urns, one for type A and one for type B.

Before the first employer in your group made a decision, we randomly (with equal probability) determined one of the two types to be the bonus type.

In the urn for the bonus type, we put **one blue** ball and **two green** balls.

In the urn for the type that is not the bonus type, we put **two blue** balls and **one green** ball.

Thus, there are **six balls in total**, three in each urn. The two urns are illustrated here.



Now, we will shuffle the balls in each urn (though every ball will remain in its own urn).

Choose a Ball

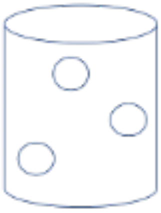
Remember that there are six balls in total, three in each urn.

We have shuffled the balls in each urn (every ball has remained in its own urn).

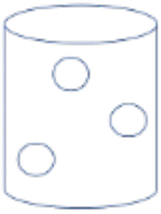
We will now show you the two urns with three balls each, but we will not reveal the ball colors.

You may then **click on exactly one ball** out of the six and we will show you its color on the next page of these instructions.

Urn for type A:



Urn for type B:



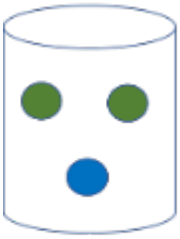
Please, select one ball to reveal its color. After you click on NEXT you will NOT be able to change your selected ball

In the following pages, we will keep showing you the two urns and the color of the ball that you revealed.

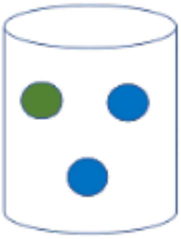
Your Choice:

Remember that the two urns are composed as follows.

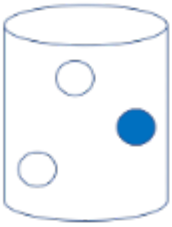
Bonus type:



Other type:

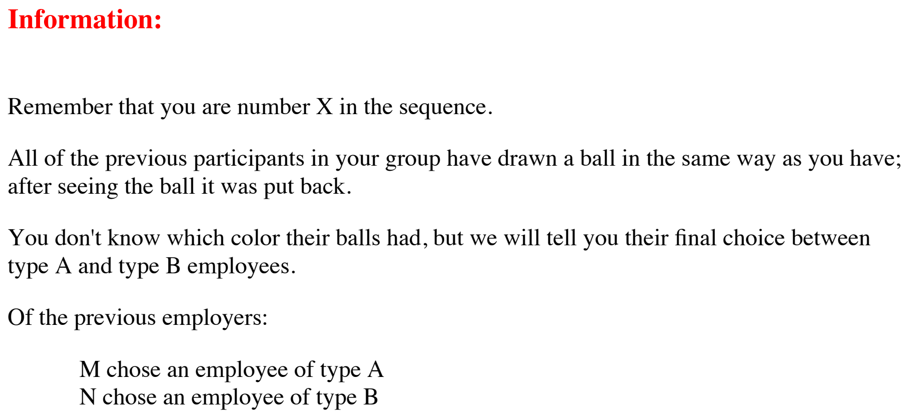


Without seeing the colors, you chose a ball from the **urn for type B**. The ball you drew is:



Now, you may use the color of the ball that you drew from the urn for type B to guess whether type B is more likely to be the bonus type or the other type.

Additional part for treatment with history information



Information:

Remember that you are number X in the sequence.

All of the previous participants in your group have drawn a ball in the same way as you have; after seeing the ball it was put back.

You don't know which color their balls had, but we will tell you their final choice between type A and type B employees.

Of the previous employers:

M chose an employee of type A
N chose an employee of type B

Your Candidates

Now, we will give you two candidates. Recall that one candidate is of type A and one candidate is of type B.

Remember that each candidate has a value between 0 and 100 points. In the end, you will receive the points of your chosen candidate plus 40 points if you choose the bonus type.

Your Choice of a Candidate

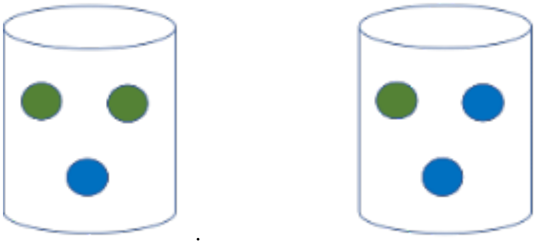
Now it's the time to choose one of the two candidates.

The one you choose will be your employee and will determine your payoff.

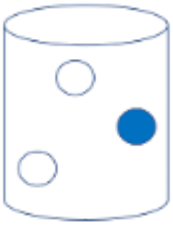
Remember:

- The two urns are composed as follows.

Bonus type: Other type:



- You chose a ball from B. The ball you drew is:



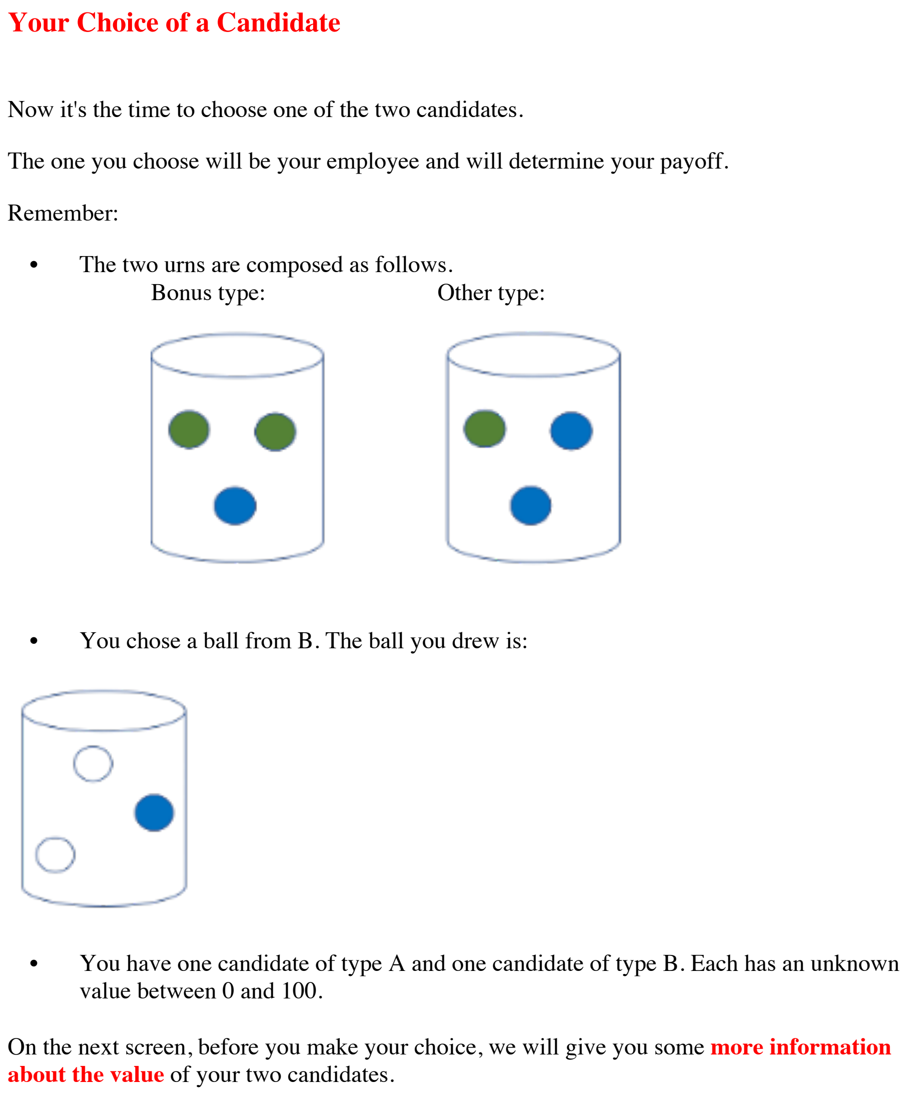
- You have one candidate of type A and one candidate of type B. Each has an unknown value between 0 and 100.

Now please choose either the candidate of type A or the candidate of type B. You will receive the value of the chosen candidate. You will receive an additional 40 points if the type you choose is the bonus type.

Please, choose your candidate:

○ A ○ B

Modified instructions for treatment with assessment software information



Your Choice of a Candidate

Now it's the time to choose one of the two candidates.

The one you choose will be your employee and will determine your payoff.

Remember:

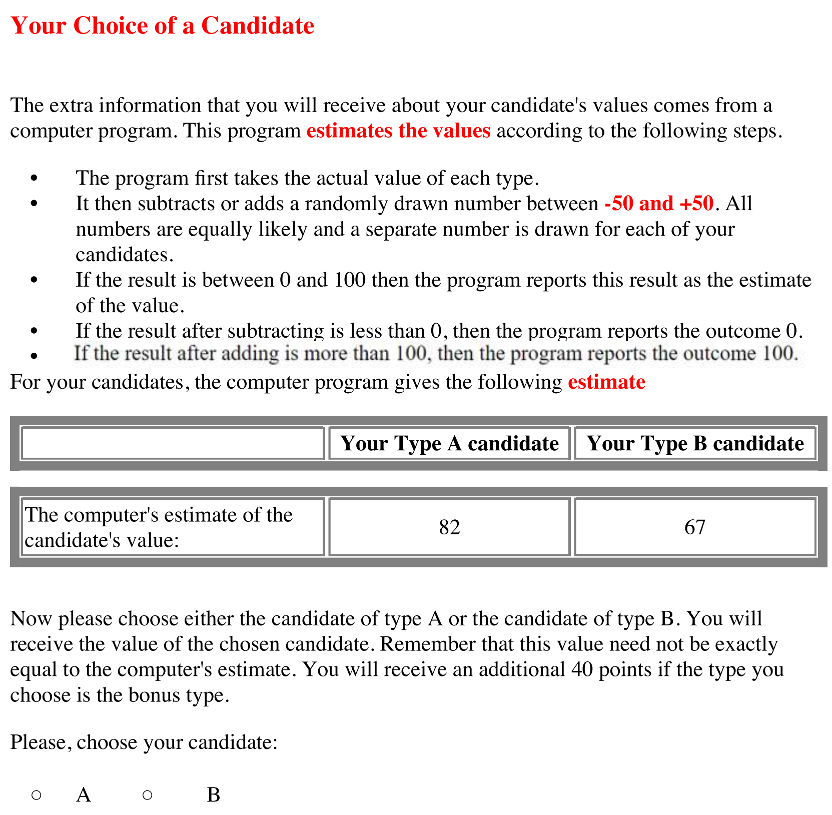
- The two urns are composed as follows.

- You chose a ball from B. The ball you drew is:

- You have one candidate of type A and one candidate of type B. Each has an unknown value between 0 and 100.

On the next screen, before you make your choice, we will give you some **more information about the value** of your two candidates.

Modified instructions for treatment with assessment software information



Your Choice of a Candidate

The extra information that you will receive about your candidate's values comes from a computer program. This program **estimates the values** according to the following steps.

- The program first takes the actual value of each type.
- It then subtracts or adds a randomly drawn number between **-50 and +50**. All numbers are equally likely and a separate number is drawn for each of your candidates.
- If the result is between 0 and 100 then the program reports this result as the estimate of the value.
- If the result after subtracting is less than 0, then the program reports the outcome 0.
- If the result after adding is more than 100, then the program reports the outcome 100.

For your candidates, the computer program gives the following **estimate**

	Your Type A candidate	**Your Type B candidate**
The computer's estimate of the candidate's value:	82	67

Now please choose either the candidate of type A or the candidate of type B. You will receive the value of the chosen candidate. Remember that this value need not be exactly equal to the computer's estimate. You will receive an additional 40 points if the type you choose is the bonus type.

Please, choose your candidate:

○ A ○ B

Your payoffs

Here are your results for today's experiment.

- Your participation fee is £ 2

- The value of the employee you chose is: X points

- As for the bonus:
    - You chose the candidate of type B
    - The randomly drawn bonus type in your group is A
    - Your total bonus is: 0

Your total earnings for participation, on top of your participation fee, are therefore X. At the exchange rate of 0.02 pounds per point this gives £ K

5

Discussion and conclusion

This chapter presents a summary of the research goals, approaches applied, major results obtained, and conclusions derived from the chapters comprising this thesis. I developed this thesis motivated by a puzzle about the use of AI technologies in hiring and their effect on gender inequality in the labor market. AI has been increasingly adopted by firms under the premise that its computational capabilities should help overcome human computational limits while ensuring unbiased decision-making processes (Black and van Esch, 2020; Langenkamp et al., 2020; Bogen and Rieke, 2018). However, existing research has shown that AI tends to learn biases from humans, reflecting existing gender biases in the labor market (Gonzalez et al., 2022; Gebru, 2020; Köchling and Wehner, 2020; Black and van Esch, 2020; Silberg and Manyika, 2019; O'neil, 2017).

While the evidence that AI may be biased is striking, the literature still encompasses some black boxes that need to be unpacked, especially considering the different AI tools firms may use in hiring and the various ways in which such tools may interact with firms' hiring decisions.

The existing literature mostly focuses on one specific type of AI, namely predictive algorithms. These algorithms predict the job candidate that firms should hire based on historical information on firms' hiring choices (Rhea et al., 2022). What we know from existing research is that predictive algorithms may discriminate against women in hiring decisions because they reflect the existing biases in historical firms' hiring choices (Gonzalez et al., 2022; Köchling and Wehner, 2020; Cowgill and Tucker, 2020). The presence of bias or underrepresentation of minorities in the data underlying algorithmic decisions is a recurrent topic in the literature documenting the potential sources of algorithmic bias (Bogen and Rieke, 2018; Daugherty et al., 2019; Silberg and Manyika, 2019; Black and van Esch, 2020; Köchling and Wehner, 2020; Gonzalez et al., 2022). To cite the most relevant study, Cowgill (2018) codified an employer's hiring choices into a dataset. He then asked an engineer to develop a predictive algorithm aimed at shortlisting job candidates for the interview based on the dataset of the employer's hiring choices. With this simple experiment, Cowgill (2018) found that not only do employers' biased hiring choices propagate from humans to the algorithm, but this is as likely to happen in the labor market as in all other settings where predictive algorithms are used for decision-making.

Yet, open questions remain on how another broadly adopted AI tool in hiring, assessment software, affects gender inequality in the labor market. Assessment software evaluates job candidates by screening their resumes or via cognitive tests, without relying on firms' past hiring choices (Li et al., 2021). Indeed, assessment software has not received the same degree of attention as predictive algorithms. To date, there is only one relevant working paper by Li et al. (2020) that studies how resume screening algorithms may affect gender and ethnic inequality in hiring. Li et al. (2020) built a screening algorithm aimed at explicitly selecting job applicants from underrepresented groups (ethnic minorities and women) to learn about their quality and potential. They found that such an algorithm improved the share of Hispanic and Black applicants shortlisted for the interview. However, they found no effect for women. Although relevant, the paper leaves the relationship between assessment software and gender obscure and does not consider how screening algorithms that are not explicitly targeted at underrepresented groups may affect inequality in hiring.

Because of the different nature of the two AI tools (predictive algorithms and assessment software), it seems reasonable to expect that predictive algorithms and assessment software may have a different effect on gender inequality in the labor market. This thesis hypothesized that predictive algorithms mimic human decision-making processes and reflect existing gender biases, as existing research suggests. This is because they use statistics to infer individual information as humans do (Gonzalez et al., 2022; Köchling and Wehner, 2020; Cowgill and Tucker, 2020). On the other hand, the thesis hypothesized that assessment software estimates the idiosyncratic quality of job applicants with higher accuracy and reliability, and reduces gender biases compared to both human recruiters and predictive algorithms. This is because they evaluate job candidates' skills through standardized tests (Williams, 2022; Raisch and Krakowski, 2021; Black and van Esch, 2020; Daugherty et al., 2019).

At this point, before moving forward with the discussion, one could well ask why this thesis deems it relevant to acknowledge whether and how predictive algorithms and assessment software differ in affecting gender inequality in hiring. Indeed, the study of gender inequality in the labor market and the relevance of the hiring process in setting the ground for gender discrimination is not new. Research shows that the hiring process plays a significant role in determining economic inequality between men and women (Reskin and Roos, 1990; Petersen and Saporta, 2004). Furthermore, experimental studies on the labor market provide evidence that during the resume screening stage of hiring, employers usually favor men, but after making it through the interview, women are just as likely as men to receive the job offer (Fernandez-Mateo and Fernandez, 2016). Since existing studies on the use of AI in hiring suggest that AI is likely to be used at the resume screening stage (cf. Bhatt (2022)), it seems reasonable to argue that AI may play a significant role in the most decisive hiring stage for gender discrimination. There is no doubt, then, about the relevance of expanding our knowledge on how the two most commonly used AI tools in hiring, assessment software and predictive algorithms, affect gender inequality in the labor market.

This thesis aimed to expand our knowledge of predictive algorithms and assessment software, starting from exploring the combined aggregate effect the two AI tools may have on gender inequality in hiring. In chapter 2, I showed, through a difference-in-differences estimate and data on Global Fortune 500 firms, that the use of AI algorithms in hiring increases by 40% - relative to the baseline - the share of female managers hired by firms. The use of AI also correlates with a reduction in firms facing gender discrimination lawsuits related to hiring.

Chapter 2 brings additional evidence, which contributes to expanding our knowledge of how predictive algorithms and assessment software affect gender inequality in the labor market. The results presented in chapter 2, although relevant, have two key limitations. First, there is still an undisclosed black box about how assessment software and predictive algorithms are used and in what settings the results presented in chapter 2 hold. Firms may use AI algorithms in two major ways: (i) they may automate the hiring process, leaving the hiring decision to AI; (ii) they may use AI to complement human recruiters and get suggestions on whom to hire, leaving the final decision to recruiters. Second, we still do not know what is the heterogeneous impact of assessment software and predictive algorithms on gender inequality in the labor market. Particularly when compared with human recruiters.

We know from chapter 2 that AI is likely to reduce gender inequality in hiring outcomes. Knowing what type of AI drives this effect and whether this is more or less likely to happen when assessment software and predictive algorithms automate the hiring process or complement human recruiters is fundamental. The question one should ask here is: how autonomous assessment software and predictive algorithms should be in making employment decisions?

Chapter 3 and chapter 4 aimed to disentangle how assessment software and predictive algorithms may affect gender inequality in the labor market when they automate hiring decisions and when they complement human recruiters, respectively.

Chapter 3 considered that gender diversity in hiring does not stand alone but interacts with the qualifications of the new hire, i.e., how likely the new worker is to meet the skills required by the job. By conducting an intervention study in one private company, I investigated whether and how assessment software and predictive algorithms that automate the hiring process affect gender diversity and the quality of the applicant pool shortlisted after resume screening. Furthermore, I compared how assessment software and predictive algorithms differ between themselves and from human recruiters in their hiring choices. I showed that although both assessment software and predictive algorithms increase to

1 the probability of hiring qualified job applicants, compared to human recruiters, only assessment software increases the probability to select female applicants, compared to human recruiters. Conversely predictive algorithms do not differ from human recruiters in the probability of selecting a more gender diverse applicant pool.

The results presented in chapter 3 provided additional evidence to the picture of how AI affects gender inequality in the labor market, thereby offering a more complete interpretation of the results presented in chapter 2. While predictive algorithms may reflect existing human biases, assessment software is likely to decrease gender inequality in hiring, particularly when it automates the hiring process. This reduction in gender inequality in hiring does not come alone but together with an increase in the qualifications of the new hire, as defined by how likely the new worker is to meet the skills required by the job.

Yet, the picture is not complete. We still need to understand what happens when assessment software and predictive algorithms complement human recruiters. In chapter 4, we modeled employers' hiring choices and how such choices are shaped by the information assessment software and predictive algorithms provide to employers. We then tested the model's predictions with an online experiment involving people with hiring experience. We found that both assessment software and predictive algorithms improve the overall productivity of selected applicants. Further, assessment software is likely to change employers' prior beliefs about job candidates and improve the diversity of the hire. This is even more pronounced when such information considerably favors one candidate over the others, making her stand out within the pool.

Indeed, the relationship between assessment software and employers' choices needs further exploration and cannot be limited to this study. However, the results presented in chapter 4 help further complete the picture of how predictive algorithms and assessment software affect gender inequality in the labor market. Considering the three empirical chapters together, the thesis not only confirmed the evidence provided by existing research about predictive algorithms reflecting human hiring choices but also argued that the effect of predictive algorithms on gender inequality in the labor market considerably differs from that of assessment software. In particular, this thesis showed that assessment software may improve the gender gap in hiring outcomes. While this is especially true when the software automates the hiring process, for the software to be effective at the intensive margin when it interacts with human recruiters, women would need to be much more qualified for the job than men. This is because the information that assessment software provides to recruiters comes from the analysis of candidates' resumes or cognitive skills. Therefore, in order for such information to be effective in changing employers' prior beliefs, women's skills and qualifications should come out as clearly higher than men's qualifications.

This last piece of evidence suggested in chapter 4 opens a wide discussion on the use of assessment software for complementing human hiring decisions. Back in the 1990s, studies already showed that employers evaluate women with stricter standards to prove competence and ability than equally qualified men (Biernat and Kobrynowicz, 1997; Foschi et al., 1994). This evidence is explained by expectation states theory (Berger, 1977; Conner, 1974), which argues that biased beliefs about gender lead to biased performance evaluation by employers and inference of ability from performance. In other words, because women are typically associated with lower social status compared to men, they need to outperform men to demonstrate their qualifications for a certain job (Ridgeway, 2011). The question here is: How different are the implications of using assessment software to complement human recruiters compared to relying solely on human recruiters? If women need to be more qualified than men for the software's suggestions to be relevant to human recruiters, is assessment software truly effective in reducing gender inequality in the labor market?

Indeed, I can empirically argue that assessment software is effective in reducing gender inequality in hiring when it complements human recruiters. Although this argument is supported by both theoretical and empirical evidence suggesting that assessment software is likely to reverse human biases, some caution is needed when making such a claim. It is hard to believe that assessment software would improve women's conditions in the labor market when it complements human decisions if women still have to prove higher qualifications than men. Indeed, as the results of chapter 4 suggest, the human-AI tandem would leave existing gender inequality in performance evaluation unaffected.

In summary, this thesis can confidently argue that, ultimately, (i) predictive algorithms have a null effect on gender inequality in the labor market; (ii) also assessment software has a null effect on gender inequality in the labor market when it is used to complement human recruiters. However, if assessment software is given full autonomy over firms' hiring decisions, it does reduce gender inequality in the labor market. This last claim brings us back to the puzzle that motivated this thesis: AI was expected to overcome human prejudices. Why then do most scholars provide evidence of AI reflecting human biases instead? The thesis answers this puzzle by arguing that AI may reproduce human biases (i) if it bases its hiring decisions on data that encode firms' past biased choices and (ii) if we restrict the autonomy of assessment software in making hiring decisions. When the human-machine tandem takes place, AI seems to be powerless in overcoming existing gender stereotypes and inequality.

From a methodological perspective, this thesis has shown that a pluralism of methods that combines traditional applied econometrics with experiments allows for a more complete and in-depth understanding of a given phenomenon. On a more personal note, I was new to experimental methods before enrolling in this Ph.D. and working with Klarita Gërxhani and Arthur Schram. Thanks to their guidance, instructions, and advice, I have learned about a completely new empirical world, which I have valued and will continue to incorporate into my research agenda.

Indeed, further research on AI and its impact on the labor market is still required. Additionally, the implications of AI for gender equality extend beyond the boundaries of labor and enter other economic domains. On a personal level, this thesis has led me to contemplate how the impact of AI tools can expand into the realm of household and development economics. While completing this thesis, I have begun to reflect on my research agenda after completing my Ph.D. and have clearly defined my research interests. Therefore, thanks to the mentoring of my Ph.D. supervisor and the collaboration with my post-doc advisors, I now know the kind of scholar I aspire to be and the high-quality research I aim to pursue. It is evident that my research interests lie in gender economics.

Indeed, the effect of AI on gender inequality will continue to hold a significant place in my research agenda. However, by maintaining a gender perspective, I am determined to pursue other research paths that aim to contribute to the literature on household and development economics.

Bibliography

Agarwal, Ritu and Vasant Dhar, “Big data, data science, and analytics: The opportunity and challenge for IS research,” 2014.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb, *Prediction machines: the simple economics of artificial intelligence*, Harvard Business Press, 2018.

Alavi, Maryam and Dorothy E Leidner, “Research commentary: Technology-mediated learning—A call for greater depth and breadth of research,” *Information systems research*, 2001, *12* (1), 1–10.

Allport, Gordon Willard, Kenneth Clark, and Thomas Pettigrew, “The nature of prejudice,” 1954.

Anderson, Lisa R and Charles A Holt, “Information cascades in the laboratory,” *The American economic review*, 1997, pp. 847–862.

Arrow, Kenneth J, “Information and economic behavior,” *HARVARD UNIV CAMBRIDGE MA*, 1973.

Banerjee, Abhijit V, “A simple model of herd behavior,” *The quarterly journal of economics*, 1992, *107* (3), 797–817.

Bhatt, Prachi, “AI adoption in the hiring process–important criteria and extent of AI adoption,” *foresight*, 2022.

Biernat, Monica and Diane Kobrynowicz, “Gender-and race-based standards of competence: lower minimum standards but higher ability standards for devalued groups.,” *Journal of personality and social psychology*, 1997, *72* (3), 544.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch, "A theory of fads, fashion, custom, and cultural change as informational cascades," *Journal of political Economy*, 1992, *100* (5), 992–1026.

Black, J Stewart and Patrick van Esch, "AI-enabled recruiting: What is it and how should a manager use it?," *Business Horizons*, 2020, *63* (2), 215–226.

Bogen, Miranda and Aaron Rieke, "Help wanted: An examination of hiring algorithms, equity, and bias," 2018.

Castilla, Emilio J, "Gender, race, and meritocracy in organizational careers," *American journal of sociology*, 2008, *113* (6), 1479–1526.

_ , "Accounting for the gap: A firm study manipulating organizational accountability and transparency in pay decisions," *Organization Science*, 2015, *26* (2), 311–333.

Chandler, Simon, "The AI chatbot will hire you now," *Wired. com*, 2017.

Cheng, Maggie M and Rick D Hackett, "A critical review of algorithms in HRM: Definition, theory, and practice," *Human Resource Management Review*, 2021, *31* (1), 100698.

Correll, Shelley J and Stephen Benard, "Biased estimators? Comparing status and statistical theories of gender discrimination," in "Advances in group processes," Emerald Group Publishing Limited, 2006.

Cowgill, Bo, "Bias and productivity in humans and algorithms: Theory and evidence from resume screening," *Columbia Business School, Columbia University*, 2018, *29.*

_ **and Catherine E Tucker**, "Algorithmic fairness and economics," *Columbia Business School Research Paper*, 2020.

Csaszar, Felipe A and Tom Steinberger, "Organizations as Artificial Intelligences: The Use of Artificial Intelligence Analogies in Organization Theory," *Academy of Management Annals*, 2022, *16* (1), 1–37.

Danieli, Oren, Andrew Hillis, and Michael Luca, "How to hire with algorithms," *Harvard Business Review*, 2016, *17.*

Datta, Amit, Anupam Datta, Jael Makagon, Deirdre K Mulligan, and Michael Carl Tschantz, "Discrimination in online advertising: A multidisciplinary inquiry," in "Conference on Fairness, Accountability and Transparency" PMLR 2018, pp. 20–34.

Daugherty, Paul R, H James Wilson, and Rumman Chowdhury, "Using artificial intelligence to promote diversity," *MIT Sloan Management Review*, 2019, *60* (2), 1.

Eagly, Alice H, "Sex differences in social behavior: comparing social role theory and evolutionary psychology.," 1997.

Ellemers, Naomi et al., "Gender stereotypes," *Annual review of psychology*, 2018, *69*, 275–298.

Fernandez-Mateo, Isabel and Roberto M Fernandez, "Bending the pipeline? Executive search and gender inequality in hiring for top management jobs," *Management Science*, 2016, *62* (12), 3636–3655.

Fernandez, Roberto M and Isabel Fernandez-Mateo, "Networks, race, and hiring," *American sociological review*, 2006, *71* (1), 42–71.

_ , Emilio J Castilla, and Paul Moore, "Social capital at work: Networks and employment at a phone center," *American journal of sociology*, 2000, *105* (5), 1288–1356.

Fiske, Susan T and Shelley E Taylor, *Social cognition*, Mcgraw-Hill Book Company, 1991.

Foschi, Martha, Larissa Lai, and Kirsten Sigerson, "Gender and double standards in the assessment of job applicants," *Social Psychology Quarterly*, 1994, pp. 326–339.

Gebru, Timnit, "Race and gender," *The Oxford handbook of ethics of aI*, 2020, pp. 251–269.

Goeree, Jacob K, Thomas R Palfrey, Brian W Rogers, and Richard D McKelvey, "Self-correcting information cascades," *The Review of Economic Studies*, 2007, *74* (3), 733–762.

Gonzalez, Manuel F, Weiwei Liu, Lei Shirase, David L Tomczak, Carmen E Lobbe, Richard Justenhoven, and Nicholas R Martin, "Allying with AI? Reactions toward human-based, AI/ML-based, and augmented hiring processes," *Computers in Human Behavior*, 2022, *130*, 107179.

Gorman, Elizabeth H, "Gender stereotypes, same-gender preferences, and organizational variation in the hiring of women: Evidence from law firms," *American Sociological Review*, 2005, *70* (4), 702–728.

Heilman, Madeline E and Tyler G Okimoto, "Why are women penalized for success at male tasks?: the implied communality deficit.," *Journal of applied psychology*, 2007, *92* (1), 81.

Hentschel, Tanja, Madeline E Heilman, and Claudia V Peus, "The multiple dimensions of gender stereotypes: A current look at men's and women's characterizations of others and themselves," *Frontiers in psychology*, 2019, *10*, 11.

Hilton, James L and William Von Hippel, "Stereotypes," *Annual review of psychology*, 1996, *47* (1), 237–271.

Houser, Kimberly A, "Can AI solve the diversity problem in the tech industry: Mitigating noise and bias in employment decision-making," *Stan. Tech. L. Rev.*, 2019, *22*, 290.

Huang, Hsieh-Hong, Jack Shih-Chieh Hsu, and Cheng-Yuan Ku, "Understanding the role of computer-mediated counter-argument in countering confirmation bias," *Decision Support Systems*, 2012, *53* (3), 438–447.

Huysman, Marleen and Dirk De Wit, "Practices of managing knowledge sharing: towards a second wave of knowledge management," *Knowledge and process management*, 2004, *11* (2), 81–92.

Jackson, Robert Max, *Destined for Equality: The Inevitable Rise of WomenÕs Status*, Harvard University Press, 1998.

Joseph, Fisek M Hamit Norman Robert Z Zelditch Jr Morris Berger, "Status Characteristics and Social Interaction," 1977.

Kahneman, Daniel, Stewart Paul Slovic, Paul Slovic, and Amos Tversky, *Judgment under uncertainty: Heuristics and biases*, Cambridge university press, 1982.

Köchling, Alina and Marius Claus Wehner, "Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development," *Business Research*, 2020, *13* (3), 795–848.

Kübler, Dorothea and Georg Weizsäcker, "Information cascades in the labor market," *Journal of Economics*, 2003, *80*, 211–229.

L, Fisek M Hamit Conner Thomas, *Expectation states theory: A theoretical research program*, Cambridge, Mass: Winthrop Publishers, 1974.

Lambrecht, Anja and Catherine Tucker, "Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads," *Management science*, 2019, *65* (7), 2966–2981.

Langenkamp, Max, Allan Costa, and Chris Cheung, "Hiring fairly in the age of algorithms," *arXiv preprint arXiv:2004.07132*, 2020.

Li, Danielle, Lindsey R Raymond, and Peter Bergman, "Hiring as exploration," Technical Report, National Bureau of Economic Research 2020.

Li, Lan, Tina Lassiter, Joohee Oh, and Min Kyung Lee, "Algorithmic hiring in practice: Recruiter and HR Professional's perspectives on AI use in hiring," in "Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society" 2021, pp. 166–176.

Lippmann, Walter, "Public opinion. 1922," *URL: http://infomotions. com/etexts/gutenberg/dirs/etext04/pbp nn10. htm*, 1965.

Oberholzer-Gee, Felix, "Nonemployment stigma as rational herding: A field experiment," *Journal of Economic Behavior & Organization*, 2008, *65* (1), 30–40.

O'neil, Cathy, *Weapons of math destruction: How big data increases inequality and threatens democracy*, Crown, 2017.

Operario, Don and Susan T Fiske, "Stereotypes: Content, structures, processes, and context," *Blackwell handbook of social psychology: Intergroup processes*, 2001, *1*, 22–44.

Petersen, Trond and Ishak Saporta, "The opportunity structure for discrimination," *American Journal of Sociology*, 2004, *109* (4), 852–901.

Phelps, Edmund S, "The statistical theory of racism and sexism," *The american economic review*, 1972, *62* (4), 659–661.

Powell, Gary N, D Anthony Butterfield, and Jane D Parent, "Gender and managerial stereotypes: have the times changed?," *Journal of management*, 2002, *28* (2), 177–193.

Prentice, Deborah A and Erica Carranza, "What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes," *Psychology of women quarterly*, 2002, *26* (4), 269–281.

Quadlin, Natasha, "The mark of a woman's record: Gender and academic performance in hiring," *American Sociological Review*, 2018, *83* (2), 331–360.

Raisch, Sebastian and Sebastian Krakowski, "Artificial intelligence and management: The automation–augmentation paradox," *Academy of Management Review*, 2021, *46* (1), 192–210.

Reskin, Barbara and Patricia A Roos, *Job queues, gender queues: Explaining women's inroads into male occupations*, Vol. 105, Temple University Press, 1990.

Reskin, Barbara F and Denise D Bielby, "A sociological perspective on gender and career outcomes," *Journal of Economic Perspectives*, 2005, *19* (1), 71–86.

Rhea, Alene K, Kelsey Markey, Lauren D'Arinzo, Hilke Schellmann, Mona Sloane, Paul Squires, Falaah Arif Khan, and Julia Stoyanovich, "An external stability audit framework to test the validity of personality prediction in AI hiring," *Data Mining and Knowledge Discovery*, 2022, *36* (6), 2153–2193.

Ridgeway, Cecilia L, "Gender, status, and leadership," *Journal of Social issues*, 2001, *57* (4), 637–655.

_ , *Framed by gender: How gender inequality persists in the modern world*, Oxford University Press, 2011.

_ , *Status: Why is it everywhere? Why does it matter?*, Russell Sage Foundation, 2019.

Rivera, Lauren A, "Pedigree," in "Pedigree," Princeton University Press, 2015.

Rudman, Laurie A and Peter Glick, "Prescriptive gender stereotypes and backlash toward agentic women," *Journal of social issues*, 2001, *57* (4), 743–762.

Sharma, A, "How AI reinvented hiring practice at L'Oréal," *People matters*, 2018, *16*, 2018.

Silberg, Jake and James Manyika, "Notes from the AI frontier: Tackling bias in AI (and in humans)," *McKinsey Global Institute*, 2019, pp. 1–6.

Simon, Herbert A et al., "Invariants of human behavior," *Annual review of psychology*, 1990, *41* (1), 1–20.

Thagard, Paul, "107C5Sociology: Prejudice and Discrimination," in "Mind-Society: From Brains to Social Sciences and Professions," Oxford University Press, 2019.

Thaler, Richard H and Cass R Sunstein, *Nudge: Improving decisions about health, wealth, and happiness*, Penguin, 2009.

Turing, Alan Mathison, "Mind," *Mind*, 1950, *59* (236), 433–460.

Vasconcelos, Marisa, Carlos Cardonha, and Bernardo Gonçalves, "Modeling epistemological principles for bias mitigation in AI systems: An illustration in hiring decisions," in "Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society" 2018, pp. 323–329.

Viale, Riccardo, *Routledge handbook of bounded rationality*, Routledge, Taylor & Francis Group, 2021.

Williams, Maryse, "Transforming Hiring and Retention: Mr. Edison Meets AI/ML," 2022.