From Transparency to Acceptance: How Explainable AI Shapes Advice Taking in Recruitment

Filipa Alegre Duarte

Dissertation written under the supervision of professor Cristina Mendonça

Dissertation submitted in partial fulfilment of requirements for the MSc in Management with a specialization in Strategy, Entrepreneurship & Impact, at the Universidade Católica Portuguesa, January 5, 2026.

Abstract

The increasing use of Artificial Intelligence (AI) in recruitment has transformed how organizations evaluate and select candidates. While AI offers efficiency and consistency, its use in high-level selection processes raises concerns about transparency, trust, fairness, and acceptance. Explainable AI (XAI) has emerged as a way to address transparency issues by providing understandable explanations for AI decisions.

This dissertation investigates whether XAI improves perceived explainability, trust, fairness, and adoption of recommendations. An experimental study was conducted in which participants received the outcome of a system regarding job recommendations.

The results indicate that XAI is substantial in perceived explainability, trust, and fairness. However, XAI does not increase participants' willingness to trust AI recommendations. Mediation analyses reveal that the variable influencing acceptance is the trust variable, not being influenced by perceived explainability. The results indicate the importance of the trust variable in associating transparency with trust.

In conclusion, this research contributed to the existing body of knowledge on AI-supported recruitment, showing that trust is still needed to promote an improved level of acceptance of AI recommendations regarding recruitment practices.

Title: From Transparency to Acceptance: How XAI Shapes Advice Taking in Recruitment

Author: Filipa Alegre Duarte

Keywords: Artificial Intelligence, Explainable AI, Recruitment, Trust in AI, Fairness, Algorithmic Decision-Making, AI Acceptance

Sumário

A crescente utilização da Inteligência Artificial (IA) no recrutamento transformou a forma como as organizações avaliam e selecionam os candidatos. Embora a IA ofereça eficiência e consistência, a sua utilização em processos de seleção de alto nível levanta preocupações sobre a transparência, confiança, imparcialidade e aceitação. A IA Explicável (IAX) surgiu como uma forma de abordar as questões da transparência, fornecendo explicações compreensíveis para as decisões da IA.

Esta dissertação investiga se a IAX melhora a explicabilidade percebida, a confiança, a justiça e a adoção de recomendações. Foi realizado um estudo experimental, no qual os participantes receberam o resultado de um sistema de em relação à recomendação de emprego.

Os resultados indicam que a IAX é substancial na explicabilidade percebida, na confiança e na justiça. No entanto, a IAX não aumenta a disponibilidade dos participantes para confiar nas recomendações da IA. As análises de mediação revelam que a variável que influencia a aceitação é a variável confiança, não sendo influenciada pela explicabilidade percebida Os resultados indicam a importância da variável confiança na associação da transparência com a confiança.

Em conclusão, esta investigação contribuiu para o corpo de conhecimento existente sobre o recrutamento apoiado por IA, mostrando que, embora seja necessária confiança para promover uma melhoria no nível de aceitação das recomendações de IA relativas às práticas de recrutamento.

Título: Da Transparência à Aceitação: Como a XAI Molda a Tomada de Conselhos no Recrutamento

Autora: Filipa Alegre Duarte

Palavras-chave: Inteligência Artificial, IA Explicável, Recrutamento, Confiança na IA, Imparcialidade, Tomada de Decisão Algorítmica, Aceitação da IA

Table of Contents

List of Figures

List of Tables

Acknowledgements

The completion of this thesis represents a very important milestone in my academic and personal journey. It marks the culmination of a demanding and enriching path, filled with challenges, learning, and growth, and stands as an achievement of which I am deeply proud.

Above all, I would like to extend my sincere appreciation to my supervisor, Professor Cristina Mendonça, for her guidance and continuous support throughout this work. Her constructive feedback, constant availability, patience, and encouragement were essential to the successful completion of this thesis. I am deeply thankful for her commitment, for consistently addressing my questions, and for providing the guidance and motivation needed to overcome the most demanding stages of this process.

I would also like to express my sincere appreciation to all the participants who took part in the survey conducted for this research. Their time, willingness, and valuable contributions were fundamental to the development and completion of this study.

Finally, I would like to acknowledge myself. As a working student, balancing professional responsibilities with academic demands was a constant challenge. This achievement reflects perseverance, discipline, and resilience, and I am proud to have successfully managed to reconcile all aspects of this journey.

I acknowledge the support from FCT – Portuguese Foundation of Science and Technology for the project UID/GES/00407/2020 and for the project 2024.07497.IACDC, supported by "RE-C05-i08.M04 – 'Apoiar o lançamento de um programa de projetos de I&D orientado para o desenvolvimento e implementação de sistemas avançados de cibersegurança, inteligência artificial e ciência de dados na administração pública, bem como de um programa de capacitação científica', of the Recovery and Resilience Plan (PRR), framed within the financing agreement entered into between the Estrutura de Missão Recuperar Portugal (EMRP) and FCT, as intermediate beneficiary.

To all those who, in one way or another, contributed to this journey, my sincere thanks.

1. Introduction

1.1 General Topic Overview

Artificial Intelligence (AI) refers to a collection of methods that enable computers to perform tasks that would typically require human reasoning and intelligence (Salin & Winston, 1992), and, although the term itself was first established in 1956 (Stuart & Norvig, 2016), AI is no longer something distant or futuristic. It already shapes how organizations make decisions every day, making an understanding of AI essential for all of us (Iansiti & Lakhani, 2020). AI systems are capable of surpassing human intelligence in various areas, particularly in structured and routine tasks that are easily automated (Parry et al., 2016; Haefner et al., 2021; Titareva, 2021), and they are widely used in modern workplaces to automate tasks and analyse large datasets. Increasingly, AI is also being applied to more complex functions - such as decision-making, creativity, and emotional intelligence - with even greater potential for these applications in the future (Kolbjørnsrud, 2017; Petrat, 2021; Kolbjørnsrud & Sannes, 2022). Rather than replacing humans, AI enhances efficiency and productivity, serving as a valuable leadership and management tool (Petrat, 2021; Wijayati et al., 2022), and a growing number of organizations are adopting it to support managerial and administrative activities (Petrat, 2021; Wijayati et al., 2022).

Technological advancements have substantially reshaped Human Resource Management (HRM) practices, enabling organizations to manage their workforce in a more effective, strategic, and professional manner (Nicolás-Agustín et al., 2022). Numerous researchers have offered varying definitions of HRM, though these perspectives largely complement one another. Schemerhorn (2001) defines HRM as the process of attracting and developing a talented workforce to support the achievement of an organization's mission, vision, goals, and objectives. Similarly, Storey (2004) views HRM as a strategic approach to employee management that aims to maintain a capable and committed workforce through cultural, structural, and personnel practices, thereby creating a competitive advantage. In practice, HRM encompasses key activities such as recruitment and selection, employee management, and training and development (Wall & Wood, 2005).

Given the broad scope and strategic importance of HRM functions, the HR department represents a significant opportunity for business process optimization (Tambe et al., 2019). In this context, the HR sector is undergoing a major transformation, often referred to as the Digital

Recruiting 3.0 era, marked by the increasing integration of AI into recruitment processes (Black & van Esch, 2019). AI is not only reshaping the roles and responsibilities of HR professionals but also profoundly influencing candidates' job application experiences (Tambe et al., 2019). As Black and van Esch (2019) argue, AI has evolved from a theoretical concept into a practical necessity within modern recruitment practices.

Furthermore, AI use in HRM processes is widespread. Amitabh & Ansari (2025) in the World Economic Forum stated that about 88% of businesses currently utilize some version of AI for preliminary candidate evaluation, while over 90% of organizations employ automated systems to sort or assess job applications. Even so, doubts remain about AI's effectiveness in recruitment (Fuller et al., 2021).

The promise sounds appealing: faster decisions, fewer human errors, and potentially less bias (Dawes, 1979; Meehl, 1954). But reality is more complex. When algorithms decide who gets an interview or how someone's potential is judged, people naturally start to ask if the selection process and outcomes were fair (Hausknecht et al., 2004). Nevertheless, risks and threats associated with AI are frequently linked to fears of machines taking over and the potential downfall of humanity (as cited by Rankila, 2020). However, Roos argues that such fears are largely unfounded though he acknowledges that the use of AI does involve certain real risks (as cited in Kemppi, 2018). This highlights the critical importance of applying ethical and responsible AI practices in recruitment and hiring processes (Tabassam et al., 2023).

Beyond technical performance, the effectiveness of AI in recruitment increasingly depends on how these systems are perceived by human decision-makers and candidates (Wesche et al., 2024). Issues such as trust (Lacroux & Martin-Lacroux, 2022), perceived fairness (Ochmann & Laumer, 2019), and acceptance (Ramachandranpillai et al., 2023) play a central role in determining whether AI-assisted hiring systems are seen as legitimate and are ultimately relied upon.

In response to these concerns, Explainable AI (XAI) has emerged, making AI models clearer and more comprehensible (Langer & König, 2022). The so-called blackbox problem - referring to AI systems whose processes and operations are hidden from the user, making it difficult to understand the rationale behind the model's decisions (Linardatos et al., 2020) - has been the focus of extensive research, particularly as AI systems are increasingly used to support critical decisions, such as healthcare interventions, criminal justice processes, and recruitment/hiring

decisions (Capone & Bertolaso, 2020). In such high-stakes contexts, it becomes essential that stakeholders are able to understand how AI systems arrive at their decisions, in order to foster transparency, trust, and their effective management (Wang et al., 2019). Moreover, explainability plays a key role in enhancing human–machine collaboration, and improving user trust (as cited in Williams, 2021). It is in response to these challenges that research into XAI has gained increasing attention. Within HRM, transparency, explainability, and interpretability are particularly critical given the highly consequential nature of the decisions involved, as they constitute key prerequisites for the sustainable adoption and use of AI systems (Janiesch et al., 2021; Mirbabaie et al., 2021). Consequently, the high-stakes character of these decisions makes HRM a domain with an inherent need for XAI.

1.2 Relevance of the Topic and Problem Statement

Due to the significant concern surrounding fairness and trust in AI, Hurlburt (2017) contended that the issue of insufficient trust warrants closer examination. Overall, there is still limited understanding of how job applicants respond to the use of algorithmic decision-making in recruitment (van Esch et al., 2019).

Trust encompasses a complex combination of interactional elements such as attitudes, beliefs, control, emotions, risk, and power (Abbass, 2019). Although trust has been widely recognized as a meaningful construct for understanding human–automation interaction, important distinctions between trust in humans and trust in automated systems must be acknowledged (Lee & See, 2004). In particular, a key challenge lies in transferring concepts of human–human trust to automation, as automated systems lack intentionality (Lee & See, 2004). This limitation is especially relevant when considering perceived purpose and intention. Trust in human relationships is strongly grounded in attributes such as loyalty, benevolence, and value alignment, which presuppose intentional mental states (Mayer et al., 1995). While automated systems do not possess genuine intentions or autonomous agency, they are developed with specific goals and therefore reflect the intentionality embedded by their designers (Rasmussen et al., 1994).

XAI has become a crucial concept in tackling these issues, helping to ensure that AI-based decision-making in recruitment is transparent, understandable, and accountable (Arrieta et al., 2020). The importance of XAI in recruitment lies in its capacity to minimize algorithmic bias, enhance fairness, and promote inclusivity within hiring practices (Hofeditz et al., 2022).

From an academic perspective, this gap - namely, the limited empirical understanding of how explainability influences trust, fairness perceptions, and behavioral acceptance of AI systems in recruitment contexts - is particularly relevant because existing research has largely focused on the technical performance and efficiency of AI systems, while offering more limited empirical insight into how explainability shapes human perceptions and behavioral responses in recruitment contexts. By examining perceived explainability, trust, fairness, and acceptance jointly, this dissertation contributes to a more comprehensive understanding of human–AI interaction in high-stakes decision-making settings.

From a managerial perspective, this research is equally relevant. As organizations increasingly rely on AI to support recruitment decisions, managers face growing pressure to ensure that these systems are not only efficient, but also perceived as fair, transparent, and legitimate. Understanding whether explainability improves trust and acceptance of AI recommendations can help organizations make more informed choices about the design and implementation of AI-based recruitment tools, thereby supporting more responsible and ethical hiring practices.

Therefore, this dissertation aims to explore whether and how explainability influences users' perceptions and acceptance of AI-assisted recruitment decisions. In other words, it examines whether providing clear and understandable explanations for algorithmic recommendations can make people feel that the system is fairer and more trustworthy. The central question guiding this research is: Does explainable AI (XAI), compared with blackbox AI, increase users' acceptance of AI recommendations in recruitment contexts by enhancing perceived explainability, trust, and fairness?

To address this broader question, the study investigates several specific aspects:

RQ1: Does XAI, compared to blackbox AI, increase trust in AI systems used for candidate evaluation?

RQ2: Does XAI, compared to blackbox AI, increase perceived explainability?

RQ3: Does XAI, compared to blackbox AI, increase perceived fairness of AI recommendations?

RQ4: Does XAI, compared to blackbox AI, lead to greater willingness to follow AI recommendations?

RQ5: Are these effects mediated by perceived explainability and trust?

1.3 Structure

To achieve the objectives of this dissertation, an experimental study was carried out. After this introductory chapter, the dissertation was organized as follows. The next chapter provided a literature review that explored key themes such as the use of AI in recruitment, the concept of XAI, trust in AI systems, and perceptions of fairness in automated decision-making. The methodology chapter then detailed the research design, sample, and the experimental procedure adopted for the study. This was followed by the results and discussion chapter, which presented the main findings and interpreted them in light of the research questions and theoretical framework. Finally, the conclusion summarized the key insights of the study, discussed their practical and theoretical implications, outlined the limitations of the research, and suggested avenues for future work on XAI in recruitment.

2. Literature Review

2.1 Artificial Intelligence in Organizational Decision-Making

AI has existed for several decades and has been applied across a wide range of fields over time. However, it is only in more recent years that the technology has advanced significantly and become widely implemented within various organizational contexts (Tecuci, 2012). Despite its long history, there is no single, universally accepted definition of AI (Legg & Hutter, 2007). One way to approach the concept of AI is by examining the individual meanings of the terms that make up the concept. The term “artificial” is generally well understood and requires little clarification (Bringsjord & Schimanski, 2003). According to the Cambridge Dictionary, “artificial” refers to something made by people, often as a copy of something natural (Cambridge University Press & Assessment, 2025). Thus, the artificial component of AI can be understood as human-made systems designed to replicate natural phenomena.

Defining intelligence, however, presents greater challenges. Some scholars describe AI as the development of machines, programs, or robots that exhibit behavior comparable to human intelligence (Tecuci, 2012; Kaplan, 2016). The difficulty with this perspective lies in the need to measure and compare human intelligence with that of machines. Kaplan (2016) offers an alternative view, describing intelligence as “the ability to make appropriate generalizations in a timely manner based on limited information” (p. 5). Other, more informal interpretations of

intelligence emphasize abilities such as reasoning, planning, learning, adapting to the environment, or retrieving and processing information (Legg & Hutter, 2007). Intelligence may also be understood as the capacity to analyze data and make decisions that take both information and context into account (Ved et al., 2016). For instance, if a system can learn to play games like tic-tac-toe, recognize faces, or generate music, it can be considered an example of AI (Kaplan, 2016). For the purposes of the current thesis, AI is defined as the capability of machines or systems to learn, interpret, and understand independently in a manner comparable to human cognitive processes.

In an organizational context, AI is considered to be a pervasive innovation that enables organizations to evolve in the digital age (Davenport, 2018b). AI has become a major driver of transformation in the field of HR, offering a wide range of possible applications (Jatobá et al., 2019). Traditional recruitment processes were revolutionized in recruitment within the past few years by AI technologies which have created unprecedented opportunities to systematize the search, evaluation, and selection process of applicants (Lundvall, 2022).

AI-powered technologies enhance the efficiency, impartiality, and predictive accuracy of identifying top talent from extensive applicant pools (Ghoneim, 2024). Such tools include résumé screening algorithms, candidate-matching systems, and automated interview platforms (Ore & Sposato, 2022).

However, these systems also have notable limitations. Key challenges involve the opaque nature of AI (i.e., the "blackbox" problem), potential biases within training data, difficulties in organizational adaptation, excessive dependence on AI-generated results, and wider ethical issues (Haenlein & Kaplan, 2019).

2.2 Blackbox AI and the Need for Explainability

Blackbox systems make it difficult to understand how decisions are made, raising serious concerns about transparency and accountability—particularly when they are deployed in high-stakes contexts (Thakur et al., 2025). The concept of blackbox AI has been widely discussed and defined in both academic and industry settings. In simple terms, it is commonly associated with a lack of understanding (Lipton, 2018) or transparency (Bucher, 2016), as the internal mechanisms and decision-making processes of these systems remain inaccessible to users.

The blackbox phenomenon refers to the difficulty of comprehending how AI systems arrive at their decisions or outputs (Marey et al., 2024). Explainable Artificial Intelligence (XAI) aims to address this challenge by enhancing the transparency of complex algorithms and shedding light on how they operate and reach decisions (Gilpin et al., 2018). Understanding the reasoning behind AI outcomes is essential for promoting algorithmic fairness, identifying biases or flaws in training data, and ensuring that systems function as intended (Gilpin et al., 2018; Marey et al., 2024).

The drive for explainability is motivated by technical, ethical, and regulatory pressures (Thakur et al., 2025), leading to an increasing demand for XAI approaches that provide meaningful insights without compromising model performance (Marey et al., 2024). Despite this progress, explanations are not yet standardized or consistently evaluated, as assessments often rely on subjective judgment (Gilpin et al., 2018; Lin et al., 2019). Consequently, the development of objective and reliable metrics for measuring the effectiveness of XAI remains an active area of research (Lin et al., 2019).

2.3 Explainable Artificial Intelligence

XAI encompasses a set of methods aimed at interpreting and understanding AI models, with a strong focus on transparency (DeSimone & Leon, 2024). Its core attributes include understandability, comprehensibility, transparency, interpretability, and explainability (Hamida et al., 2024). Although these terms are often used interchangeably, important distinctions exist - particularly between explainability and interpretability (DeSimone & Leon, 2024). Explainability refers to explanations functioning as an interface between users and AI systems (Angelov et al., 2021) aiming to clarify internal decision processes and justify model outputs, even in complex and opaque models such as deep neural networks (Johnson, 2020; Band et al., 2023). In contrast, interpretability concerns the extent to which a system can directly link causes to effects through inherently understandable mechanisms (Johnson, 2020), as well as the degree to which users can comprehend these explanations without additional layers (Markus, 2021).

The main objective of XAI is to help humans better comprehend AI systems, foster informed trust, and support the development of models that are more transparent and explainable (Dwivedi et al., 2023). As reliance on intelligent systems increases, the demand for interpretable and transparent models continues to grow, especially in high-stakes domains where explainability is becoming a benchmark for trust and deployment (Dwivedi et al., 2023). Clear

explanations of automated decisions contribute to demonstrating reliability, building user confidence, validating predictions, and identifying hidden errors or biases, thereby encouraging more responsible AI use (Ortigossa et al., 2024).

In contrast to traditional blackbox models, XAI provides human-understandable explanations for recommendations and outcomes by clarifying the reasoning and criteria behind algorithmic decisions (Ghoneim, 2024). This enables stakeholders - such as recruiters, hiring managers, and job applicants - to more easily interpret and evaluate AI-generated decisions (Shin, 2021). By emphasizing transparency alongside accuracy and fairness, XAI plays a crucial role in strengthening organizational trust as AI systems become increasingly embedded in decision-making processes (Hamida et al., 2024).

2.4 XAI in Recruitment and HRM

Human Resources (HR) practices are increasingly incorporating AI for tasks such as recruitment, performance evaluation, and employee retention (Gopi et al., 2025). Despite this growing adoption, the need for AI systems in HR that are transparent and easy to interpret remains largely unmet (Gopi et al., 2025).

Applying XAI techniques to recruitment data aims to uncover the key factors driving hiring decisions and to enhance transparency in the recruitment process (Shulner-Tal & Sheidin, 2025). However, despite their potential, these techniques often struggle to effectively integrate their outputs into visualizations and user interfaces that are accessible to non-technical HR users (Gopi et al., 2025).

The relevance of XAI in recruitment lies in its capacity to mitigate algorithmic bias, promote fairness, and foster inclusivity in hiring practices (Hofeditz et al., 2022). Although AI systems are frequently perceived as objective, human biases can unintentionally influence decision rules, training datasets, and algorithmic structures, causing AI to reproduce or even amplify existing inequalities throughout the recruitment process (Hofeditz et al., 2022). Empirical evidence clearly demonstrates that discrimination in hiring is well documented, particularly concerning age, gender, and race (Baert, 2018). This body of evidence suggests that AI systems trained on historical recruitment data cannot be assumed to be inherently objective and may, without targeted interventions, perpetuate established patterns of discrimination. In this context, XAI does not eliminate bias but plays a crucial role in making discriminatory decision patterns

visible and open to scrutiny, thereby supporting more accountable governance of AI-driven recruitment systems (Hofeditz et al., 2022).

Beyond bias mitigation, XAI also strengthens stakeholder trust and acceptance of AI technologies by providing meaningful insights into how algorithmic decisions are made (Hofeditz et al., 2022). For recruiters and hiring managers, transparent explanations support more informed and accountable decision-making, enabling them to assess the relevance, accuracy, and fairness of AI-generated recommendations (Rigotti & Fosch-Villaronga, 2024). Similarly, job applicants benefit from increased transparency by gaining a clearer understanding of the factors influencing their evaluations, which enhances trust in the system and perceptions of fairness within the recruitment process (Sheridan et al., 2022).

The integration of XAI into recruitment thus represents a significant advancement in digital service innovation. It enables organizations to respond to evolving workforce needs through more efficient hiring processes, improved understanding of candidate profiles, and data-driven decision-making supported by AI algorithms (Tambe et al., 2019). Moreover, a focus on ethical decision-making and stakeholder trust aligns with the core principles of responsible innovation in the digital services sector (Ghoneim, 2024). Through XAI, organizations can enhance HR practices, uphold ethical standards, and strengthen stakeholder trust, thereby advancing digital innovation while mitigating bias, promoting diversity, and improving overall organizational performance, competitiveness, and sustainability (Arrieta et al., 2020).

2.5 Trust in AI Systems and Perceived Explainability

2.5.1 Trust in AI Systems

Although the theoretical advantages of XAI are persuasive, its real-world implementation provides even deeper insights. A prime example of this is LinkedIn's recruitment tools, which employ AI to improve job matching while emphasizing transparency (Ghoneim, 2024). As the world's largest professional networking platform, LinkedIn leads the way in applying AI to recruitment (Ghoneim, 2024). Its sophisticated algorithms align job seekers with suitable opportunities by analyzing factors such as location, industry trends, and professional experience (Ghoneim, 2024). A key component of LinkedIn's AI approach is the integration of XAI, which increases the transparency and interpretability of its decision-making processes (Ghoneim, 2024). Through XAI, LinkedIn enables users to understand the reasoning behind job

recommendations, thereby building trust by making the AI's suggestions more transparent and explainable (Nechytailo, 2023; LinkedIn Engineering, 2021).

Empirical research suggests that XAI has the potential to enhance users' trust in AI systems; however, this relationship is complex and contingent on multiple factors (Weitz et al., 2019; De Brito Duarte et al., 2023; Rosenbacke et al., 2024). Although explanations often foster greater trust - particularly when they are clear, concise, and relevant - their presence alone does not guarantee increased trust, and in certain contexts, XAI may even have a negative effect (De Brito Duarte et al., 2023; Rosenbacke et al., 2024). The quality of explanations is therefore crucial, as overly complex or inconsistent explanations can erode user confidence (Rosenbacke et al., 2024). Moreover, system performance frequently exerts a stronger influence on trust than explainability alone (De Brito Duarte et al., 2023). Beyond trust in AI systems themselves, XAI also shapes trust in AI developers and engineers, with perceived explainability playing a significant role in shaping trust-related perceptions, which subsequently influence public attitudes and intentions to adopt AI systems (Cheung & Ho, 2025). Importantly, trust is a multidimensional construct: while explanations contribute cognitive information that supports informed trust judgments, affective factors and information beyond explanatory depth also moderate the formation of trust (Bernardo & Seva, 2023).

In sum, although the effects of XAI on trust are not uniform, explainability offers advantages over blackbox AI in supporting users' trust judgments. Thus, I propose the following:

> H1: Participants exposed to XAI will report higher trust in the AI system than those exposed to blackbox AI.

2.5.2 Perceived Explainability

Explainability in AI involves understanding the underlying logic and processes that lead to an algorithm's specific outcomes and decisions (Arrieta et al., 2020). A related concept, causability, refers to how effectively an explanation enables users to grasp the core reasoning behind a given result (Holzinger et al., 2019). When AI decisions lack interpretability, this opacity can reinforce existing biases and potentially result in unjust or discriminatory outcomes (Quinn et al., 2022).

Although XAI aims to foster AI adoption by increasing algorithmic transparency, its direct impact on users' perceived explainability is often modest (Hamm et al., 2023). Notably,

perceived explainability itself strongly influences both trust and perceived usefulness (Hamm et al., 2023). Users may even accept XAI systems with lower accuracy when explanations are relevant and evidence-based, suggesting a preference for rich contextual insight over simple outcome reporting (Rastogi et al., 2025).

Interactive XAI approaches can further enhance perceived usefulness and team performance, though they may also increase task completion time and yield mixed effects on cognitive load and overconfidence (Bertrand et al., 2023). Paradoxically, some studies indicate that reduced explainability can sometimes lead to higher perceived trustworthiness and competence (Hauptman et al., 2024). Moreover, post-hoc explanations of blackbox models may be incomplete or biased, underscoring the importance of behavioral measures for a more holistic understanding of interpretability (John-Mathews, 2021).

Taken together, while prior research indicates that the effects of XAI on perceived explainability are neither uniform nor guaranteed, the inclusion of explanation mechanisms is nevertheless expected to increase users' perceived explainability relative to fully opaque systems. Accordingly, the following hypothesis is proposed:

> H2: Participants exposed to XAI will report higher perceived explainability compared to those exposed to blackbox AI systems.

2.6 Fairness and Acceptance of AI-Based Decisions

2.6.1 Perceived Fairness in AI-Assisted Recruitment

Perceived fairness reflects how people judge the AI-driven recruitment process in terms of equity and openness (Singh & Finn, 2022). It encompasses their beliefs about whether the system is ethical, unbiased, and treats all applicants equally (Singh & Finn, 2022). Fairness and freedom from bias are considered key ethical priorities in recruitment (Hofeditz et al., 2022).

Applicants often receive little insight into how algorithms evaluate their qualifications or determine their rankings and this lack of clarity can undermine trust and perceived fairness, ultimately reducing candidates' willingness to accept AI-driven hiring processes (Chamorro-Premuzic et al., 2023).

Research suggests that although XAI has been widely studied for its ability to reveal and interpret unfairness (Papanikou et al., 2025), as well as to mitigate and assess bias

(Ramachandranpillai et al., 2023), its direct empirical effects on users' fairness perceptions - particularly in comparison to blackbox models - remain underexplored. Some studies indicate that explanations can improve transparency and perceived trustworthiness (Ramachandranpillai et al., 2023), and that XAI techniques may assist end users in identifying and potentially correcting fairness-related issues in AI systems (Nakao et al., 2022). Nevertheless, many assertions about XAI's fairness advantages lack clear definitions and normative foundations (Deck et al. 2024), and XAI should therefore be regarded as only one of several tools for promoting algorithmic fairness (Deck et al. 2024). A critical survey emphasizes the importance of clearly specifying the XAI approach used, the targeted fairness objective, the mechanism through which fairness is supported, and the stakeholder who benefits (Deck et al. 2024). Moreover, evaluations of XAI systems must explicitly account for fairness, as variations in user backgrounds can introduce biased perceptions and influence AI-driven outcomes (Nguyen et al., 2024). To address these challenges, a proposed evaluation framework integrates fairness-related properties and metrics to better ensure equitable results (Nguyen et al., 2024).

Overall, although XAI is not a standalone solution for algorithmic fairness, prior research suggests that explainability can enhance transparency and support users' fairness judgments compared to opaque, blackbox systems. Thus:

> H3: Participants exposed to XAI will perceive the system as fairer compared to those exposed to blackbox AI.

2.6.2 Acceptance of AI Recommendations

Although XAI has the potential to improve semantic interpretation in résumé screening, increase accuracy in job matching, and support more thorough analysis in interview assessments (Zhang et al., 2025), its effectiveness depends heavily on contextual conditions and users' levels of AI literacy (Fleiß et al., 2024; Kalff & Simbeck, 2025). For example, the specific types of skills being evaluated can shape how explanations influence user acceptance (Fleiß et al., 2024), and HR managers' AI literacy plays a significant role in how well XAI components are understood (Kalff & Simbeck, 2025). Users with moderate to high AI literacy tend to report more positive subjective perceptions of XAI, though this does not always translate into better objective understanding (Kalff & Simbeck, 2025). Moreover, many existing recruitment dashboards provide insufficient explanations of AI-generated outcomes (Ahmed, 2024). As a result, enhancing AI literacy and adapting explanation strategies to user needs are

essential for ensuring the fair and effective integration of AI in recruitment processes (Zhang et al., 2025).

Taken together, these findings suggest that, while the effectiveness of XAI depends on contextual factors and users' AI literacy, explainable systems can facilitate understanding and acceptance of AI outputs. Compared to opaque, blackbox systems, this increased transparency is expected to strengthen users' willingness to follow AI recommendations. Hence, I propose the following:

> H4: Participants in the XAI condition will show greater willingness to follow the AI's recommendation compared to those exposed to blackbox AI.

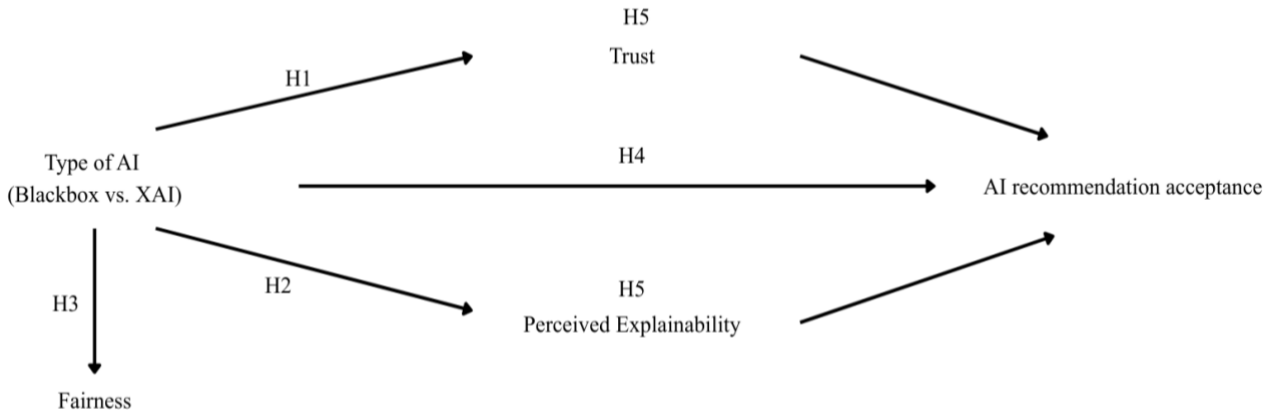
As discussed in Section 2.5, perceived explainability and trust play central roles in shaping how users interpret, evaluate, and rely on AI outputs. When users understand the rationale behind AI recommendations and perceive the system as trustworthy, they are more likely to accept and act upon its advice. Thus:

> H5: The effect of AI type (XAI vs. blackbox) on participants' willingness to follow AI advice will be mediated by perceived explainability and trust.

2.7 Conceptual Framework and Hypotheses

Figure 1 presents the conceptual model of this study, which proposes that the type of AI system - blackbox vs. XAI - influences how users understand and respond to AI-generated recommendations in recruitment contexts. Trust and perceived explainability are modeled as parallel mediators of the relationship between AI type and willingness to follow AI advice (H5), while fairness is examined as a distinct, secondary outcome of AI type.

Figure 1 - Conceptual framework and hypotheses



3. Methodology

3.1 Research Design

This dissertation examines whether increasing the transparency of AI systems influences participants' perceived explainability, trust, fairness, and behavioural acceptance in AI-assisted hiring decisions. The aim is to empirically assess whether providing insight into how an AI system evaluates candidates makes the system feel more understandable, trustworthy, and procedurally fair to users, and whether these perceptions shape their willingness to follow the AI's recommendation. The contribution of this study lies in examining explainability not as an abstract technical property, but as a mechanism that may meaningfully improve human responses to AI-assisted recruitment.

To investigate these questions, a 2 × 2 between-subjects experimental design was implemented in Qualtrics. Two elements were manipulated: 1) the type of AI system providing the hiring recommendation - XAI or Blackbox AI - and 2) the order in which the two candidate profiles were presented (stronger candidate first vs. weaker candidate first). These manipulations resulted in four experimental conditions, to which participants were assigned.

Random assignment was used to minimize systematic differences between groups in variables such as prior experience with AI, demographic characteristics or attitudes toward technology. Although randomization does not guarantee perfect equivalence across conditions, it is a method for reducing potential confounds and increasing internal validity by ensuring that

observed differences are more likely attributable to the manipulated variables (Cook et al., 2002).

A between-subjects design was intentionally chosen so that each participant viewed only one scenario. This decision was important to prevent participants from becoming aware of the experimental manipulation by comparing different types of AI explanations or different presentation orders. It also avoided learning or contrast effects that could arise if participants were exposed to multiple scenarios (Charness et al., 2012). By restricting each participant to a single condition, the design strengthened the interpretability of differences across groups and ensured that responses more closely reflected participants' natural reactions to a single AI-assisted hiring decision.

Together, this experimental approach provided a controlled way to examine whether explainable AI can meaningfully enhance perceived explainability, trust, and fairness, and whether these perceptions translate into greater behavioural reliance on AI recommendations in a recruitment context.

3.2 Participants

The study was distributed through multiple online channels, which featured a website focused on finding research participants (SurveyCircle), my own social media (Instagram, LinkedIn, Facebook, X), as well as my personal networks. Additionally, it is important to note that every participant in this research was a volunteer.

Of the 204 responses collected, 185 successfully passed a manipulation check. Following this, an additional exclusion criterion based on completion time was applied. Given that the main textual content of the study comprised approximately 700 words and that the average reading speed is estimated at around 0.25 seconds per word (Brysbaert, 2019), participants would require a minimum of approximately 175 seconds to read the material alone, not accounting for page transitions or response selection. Therefore, participants who completed the study in less than 175 seconds were excluded, as such response times likely indicate insufficient engagement with the study materials. This resulted in the exclusion of 37 additional participants (one participant had already been excluded for failing the attention check), yielding a final sample of 148 participants.

This final sample size is considered adequate, as prior research indicates that parallel mediation models with two mediators using bootstrapping techniques require approximately 110 to 130 participants to reliably detect medium-sized effects (Sim et al., 2022).

Among these participants, 58% identified as female and 42% as male, with an average age of 30.06 years (*SD* = 12.12). Most respondents were highly educated, with 90,5% holding at least a bachelor's degree, and the majority (86%) reported living in Portugal.

Participants also varied in their familiarity with AI tools: 32% considered themselves to have low experience, while 58% reported a medium level of experience. In terms of professional background, 62% worked in management or administration, and 8% in human resources or recruitment.

3.3 Procedure

Participants began the study by reading an introductory page that included a brief description of the research and explained that continuing to the next page constituted informed consent to participate. After providing their consent, participants were asked to indicate the name of a company for which they either currently worked or would like to work; this response was used only to personalize the recruitment scenario and enhance its realism.

Next, participants were introduced to the evaluation task: they were asked to put themselves in the role of an HR manager who has to choose a candidate for a Data Scientist position. They were then shown two candidate profiles. The order in which these profiles appeared - stronger candidate first or weaker candidate first - was determined through random assignment programmed in Qualtrics.

Following the review of both candidates, participants indicated their initial hiring preference using a continuous slider ranging from 0 ("strongly prefer Sam Jones") to 100 ("strongly prefer Jess Smith"). This pre-recommendation measure served as a baseline indicator of participants' preferences prior to any AI input.

Participants were then randomly allocated to either of the two AI experimental conditions: blackbox AI vs XAI. In the blackbox condition, the AI system provided a screening score and hiring recommendation with no explanation regarding its reasoning. In the XAI condition, the same recommendation was accompanied by a short, clear explanation highlighting the criteria

that informed the AI's decision. As an example (note the difference between blackbox and AI was the omission, in the blackbox condition, of the AI Explanation part):

> Please imagine the company [company name] uses an AI decision-support system that analyzes candidate profiles and provides a recommendation score on a 0-100 scale, where, in this case, 0 = Strongly prefers Sam Jones and 100 = Strongly prefers Jess Smith. For these two candidates, the output of the AI system was:
> AI Recommendation Score: 25;
> AI Recommendation: Sam Jones;
> AI Explanation: Sam Jones scored higher due to their graduate education in Data Science, four years of relevant experience, and advanced technical skills (Python, Power BI, TensorFlow) that align with the job requirements. Jess Smith scored lower because their education and experience are more limited, and their technical background is less specialized for this role.

After viewing the AI recommendation (with or without explanation), participants once again indicated their hiring preference on the same 0-100 slider. This second preference measure allowed for assessment of behavioural reliance on the AI recommendation.

The following block contained the study's key measures. Participants first completed the perceived explainability, perceived fairness of the AI-assisted hiring process, and trust measures. Then, a manipulation check asked participants whether the AI system they were exposed to provided an explanation.

Before the demographics, which concluded the survey, there was an opportunity for participants to provide open-ended comments or feedback related to the scenario, the AI system, or the questionnaire in general. This section allowed the respondents to elaborate on their experience in their own words and provided qualitative insights that could supplement the quantitative results.

The last questions in the questionnaire were on demographic characteristics - age, gender, level of education, country of residence, professional background, and experience with AI systems (which served as control variables) - and ended with a 'thank you' note and the contact information of the researcher, in case participants wanted to get more information on this study. A full version of the questionnaire is given in Appendix A.

3.4 Variable Measurement

3.4.1 Dependent Variables

Fairness: Perceptions of fairness in the AI-assisted hiring procedure were measured using a three-item scale adapted from Schoeffer et al. (2024). These items evaluate the perceived ethicality and impartiality of the AI system (e.g., original item: "The AI's procedures are free of bias"; adapted item: "The AI system used for candidate selection is free of bias"). All items were rated on a 5-point Likert scale (1 = "strongly agree"; 5 = "strongly disagree").

AI recommendation acceptance: Participants' reliance on AI recommendation was measured using the Weight of Advice (WOA) index (Yaniv, 2004), which captures the extent to which participants adjusted their hiring decision toward the AI's recommendation. WOA was calculated as the proportional change from participants' initial preference to their final preference, relative to the distance between the initial preference and the AI's advice. Preferences were recorded using a continuous slider ranging from 0 ("strongly prefer Sam Jones") to 100 ("strongly prefer Jess Smith"), with the AI recommendation expressed on the same scale. The AI output was either 25 (in the condition where Sam was the stronger candidate, indicating a recommendation for Sam) or 75 (in the condition where Jess was the stronger candidate, indicating a recommendation for Jess). Cases in which participants' initial preference coincided with the AI recommendation ($n = 10$) were excluded from the computation, as advice taking is not meaningful when no adjustment toward the advice is possible. Some WOA values fell outside the theoretically meaningful range of −1 to 1. Following prior advice-taking research (e.g., Gino & Moore, 2007), these extreme values were addressed through winsorization by recoding values above 1 to 1 and values below −1 to −1. All analyses involving AI recommendation acceptance were conducted using the winsorized WOA measure.

3.4.2 Independent Variable

AI System Type: The primary independent variable in this study was the kind of AI system that offered the hiring suggestion. Participants were allocated to either the XAI group or the blackbox AI group. As described in the literature review, the main distinction between these systems is the level of transparency they provide. While both systems delivered a hiring recommendation accompanied by an AI-generated score, only the XAI condition included an explanation describing the criteria the AI relied upon (e.g., education, experience, and technical skills). In contrast, the blackbox condition presented the recommendation without any insight

into the AI's reasoning. Aside from the presence or absence of the explanation, all other content presented to participants was held constant across conditions. A dummy variable was created to represent the main experimental manipulation (0 = blackbox AI, 1 = XAI).

3.4.3 Mediator Variables

Trust: Trust toward the AI-based hiring system was assessed using a five-item adaptation of the Trust in Automation scale originally developed by Merritt (2011). This scale has been widely applied in automation and AI-acceptance research. The items were adjusted to match the recruitment context (e.g., original item: "I usually trust machines until there is a reason not to"; adapted item: "I usually trust AI systems in recruitment decisions until there is a reason not to"). One item was reverse-scored ("For the most part, I distrust AI systems used in hiring decisions."). All items were measured on a 5-point Likert scale from "strongly agree" to "strongly disagree".

Perceived explainability: Participants' perceptions of how clear and understandable the AI system's decision-making process was were measured using an adapted version of the Transparency Of RObots (TOROS) scale (Angelopoulos et al., 2025). This scale has been used in recent studies examining transparency in intelligent systems. The adapted version contained four items (e.g., original item: "I feel like the robot's explanations are useful"; adapted item: "The AI system's explanations are useful"), each rated on a 5-point Likert scale ranging from 1 = "strongly agree" to 5 = "strongly disagree".

3.4.4 Control Variables

Experience with AI tools: Previous research shows that past experience consistently shapes how people interact with automated systems, independent of the specific direction of its effect (Hoff & Bashir, 2015). Experience matters because it helps users build a clearer understanding of how an automated system works and what it is intended to do (Hoff & Bashir, 2015). For this reason, participants in the present study were asked to report their level of experience with AI tools using a single-item measure ("What is your experience with AI tools?"), rated on a four-point scale (1 = "none", 2 = "low", 3 = "medium", 4 = "high"). This measure was included as a control variable to account for individual differences in familiarity and exposure to AI systems.

Demographic Controls: Demographic characteristics were also collected and considered as potential control variables. These were participants' age, gender, education level, professional background, and country of residence. Country of residence was dummy coded, with Portugal coded as 1 and all other countries coded as 0. Education level was treated as an ordinal variable (1 = High School, 2 = Bachelor's degree, 3 = Master's degree, 4 = PhD). Professional background was dummy coded into four categories: Human Resources/Recruitment (1), Management/Administration (2), Other professional backgrounds (3), and no professional background (4).

4. Results

4.1 Scale Reliability

Internal consistency was assessed for all multi-item scales using Cronbach's α, adopting the commonly accepted threshold of $\alpha \geq .70$ for adequate reliability (Bland & Altman, 1997). As shown in Table 1, all scales met this criterion except the fairness scale ($\alpha = .68$). Although adapted from Schoeffer et al. (2024) and grounded in procedural justice frameworks (Colquitt & Rodell, 2015), this three-item scale fell below the recommended threshold, which is a common limitation of shorter scales (Schweizer, 2011).

Table 1- *Scale's reliability test results*

Scale	Number of items	Cronbach's α
Trust	6	.87
Explainability	4	.93
Fairness	3	.68

4.2 Descriptive Statistics

The final sample consisted of 148 participants. Of these, 66 participants were assigned to the blackbox AI condition, while 82 participants were assigned to the XAI condition. Descriptive statistics for all scale variables examined in the experiment are reported in aggregated form in Table 2 and are further examined in the hypothesis testing section.

Table 2 - *Descriptive statistics of all the variables used*

Scale	*N*	Min.	Max.	*M*	*SD*
Trust	148	1	5	3.23	0.73
Explainability	148	1.25	5	3.65	1.05
Fairness	148	1.67	5	3.70	0.69
Pre-AI Preference	148	0	100	54.16	35.29
Post-AI Preference	148	0	100	51.40	36.47
AI Recommendation Acceptance	138	-1	1	0.37	0.57

4.3 Bivariate Correlations

As an exploratory and initial analysis, Pearson bivariate correlations were conducted to examine the relationships among the key variables of the conceptual model, in order to better understand their associations and to provide a preliminary assessment of hypotheses H1 to H4.

Additionally, it is possible to investigate associations between types of AI systems (blackbox or XAI), perceptions regarding explainability, trust, perception of fairness, and acceptance of AI system recommendation for further statistical testing.

As illustrated in Table 3, system type was positively and significantly related to trust ($r = .22$, $p = .006$). This confirms the hypothesized prediction provided in H1 since individuals who used the XAI system (coded as 1) reported higher levels of trust towards the AI system than those who used the blackbox system (coded as 0).

There was a strong positive relationship between the type of AI system and the perceived explainability ($r = .66$, $p < .001$), which provided strong evidence in support of H2. The correlation coefficient suggests that the experimental manipulation was highly successful in allowing the two conditions of the AI system to differ on perceived transparency.

The type of AI system was also positively and significantly correlated with fairness perceptions ($r = .26$, $p = .001$), lending initial support for H3. The findings indicate that exposure to XAI is

associated with higher perceptions of fairness in the AI-assisted hiring decision. In contrast, the correlation between type of the AI systems and acceptance of the AI recommendation was positive but did not reach statistical significance (r = .15 and p = .080). Therefore, H3 was not supported at the bivariate level. This finding suggests that, although participants exposed to XAI tended to adjust their hiring preferences slightly more in the direction of the AI recommendation, this direct association was not strong enough to be statistically reliable.

Table 3 - *Bivariate correlations among study variables*

Variable	N	1	2	3	4	5	6	7	8	9	10	11
1. Trust	148	---										
2. Explainability	148	.39**	---									
3. Fairness	148	.47**	.32**	---								
4. AI Advice Acceptance	138	.21*	.15	.40**	---							
5. AI System Type	148	.22**	.66**	.26**	.15	---						
6. Age	148	.06	-.02	.08	-.09	-.01	---					
7. Male	148	.10	-.03	.22**	.09	.02	.14	---				
8. Portugal Resident	148	-.12	.06	-.08	-.17*	.00	.08	.02	---			
9. Experience with AI Tools	148	.03	.12	-.31**	-.09	-.11	-.27**	-.05	.05	---		
10. Education Level	148	.11	-.00	.02	.14	-.02	-.30**	-.05	-.05	.27**	---	
11. Professional Background	148	-.20*	-.03	-.14	-.15	-.03	-.01	.17*	.02	.03	.27**	---

Note: **. Correlation is significant at the 0.01 level (2-tailed); *. Correlation is significant at the 0.05 level (2-tailed).

4.4 Hypothesis Testing

4.4.1 The Effect of AI System Type on Perceived Fairness

H3 stated that XAI would lead participants to perceive the system as fairer compared with blackbox AI. To test this hypothesis, a multiple regression analysis was conducted with perceived fairness as the dependent variable. AI system type was included as the main independent variable, together with a set of control variables comprising age, gender, education level, professional background, country of residence, and experience with AI tools.

The overall regression model was statistically significant, $F(7, 140) = 5.95$, $p < .001$, explaining 22.9% of the variance in perceived fairness ($R^2 = .23$; adjusted $R^2 = .19$). Importantly, AI system type emerged as a significant positive predictor of fairness perceptions ($b = 0.31$, $SE = 0.10$, $\beta = 0.22$, $p = .003$). This result indicates that participants exposed to the explainable AI system

perceived the recruitment decision-making process as significantly fairer than those exposed to a blackbox AI system, even after accounting for individual differences.

Regarding the control variables, both gender and experience with AI tools were significant predictors of perceived fairness. Male participants reported higher fairness perceptions compared to female participants. In contrast, experience with AI tools was negatively associated with perceived fairness, suggesting that participants with greater familiarity with AI technologies tended to evaluate the system more critically. Detailed regression results are provided in Appendix B.

Crucially, the effect of AI system type remained significant after controlling for all demographic and experiential factors, indicating that the positive influence of explainability on fairness perceptions is robust. Overall, these findings provide clear support for H3.

4.4.2 Mediation Analysis: The Role of Trust and Perceived Explainability

To jointly test H1, H2, H4, and H5, a parallel mediation analysis was conducted using PROCESS Model 4 (Hayes, 2022). AI system type (0 = blackbox AI; 1 = XAI) was specified as the independent variable, perceived explainability and trust as parallel mediators, and AI recommendation acceptance as the dependent variable. The analysis was based on 5,000 bootstrap resamples with 95% confidence intervals.

Consistent with H1, AI system type significantly predicted trust in the system. Participants exposed to the XAI condition reported higher trust than those exposed to the blackbox AI, with a positive and statistically significant effect ($b = 0.31$, $SE = 0.12$, $t = 2.48$, $p = .014$). Likewise, and in line with H2, AI system type had a strong and significant effect on perceived explainability. Exposure to XAI substantially increased perceived explainability compared to blackbox AI ($b = 1.42$, $SE = 0.13$, $t = 10.71$, $p < .001$). These results confirm that the explainable system successfully enhanced both users' understanding of the AI's decision-making process and their confidence in the system.

Turning to behavioral outcomes, the total effect of AI system type on AI recommendation acceptance was positive but only marginally significant ($b = 0.05$, $SE = 0.03$, $t = 1.75$, $p = .080$). As such, H4, which predicted greater willingness to follow AI recommendations in the XAI condition, was not clearly supported at the total effect level. When perceived explainability and trust were included in the model, the direct effect of AI system type on recommendation

acceptance became non-significant ($b = 0.03$, $SE = 0.03$, $t = 0.89$, $p = .375$), suggesting that the influence of XAI on behavioral acceptance is not direct.

The mediation analysis revealed a differentiated pattern across the two mediators. The indirect effect via trust was statistically significant (indirect effect = 0.04), with a 95% bootstrap confidence interval that did not include zero (BootLLCI = 0.003; BootULCI = 0.092). This indicates that XAI increased AI recommendation acceptance indirectly by increasing trust in the system. In contrast, the indirect effect via perceived explainability was not significant (indirect effect = 0.012), as the corresponding bootstrap confidence interval included zero (BootLLCI = −0.156; BootULCI = 0.186). Thus, while XAI strongly increased perceived explainability, this increase did not translate into greater behavioral acceptance of the AI's recommendation.

Overall, these findings provide partial support for H5. The results demonstrate that the effect of AI system type on willingness to follow AI advice is mediated by trust, but not by perceived explainability. Although explainability plays a crucial role in shaping users' cognitive evaluations of the system, trust emerges as the key psychological mechanism through which transparency is converted into behavioral acceptance. The parallel mediation analysis was re-estimated including age, gender, education level, professional background, country of residence, and experience with AI tools as covariates. The inclusion of control variables did not alter the pattern of results: trust remained the only significant mediator, whereas the indirect effect via perceived explainability remained non-significant. The full PROCESS output is reported in Appendix C.

5. Discussion

5.1 Summary of Results

The present dissertation set out to examine whether increasing the transparency of AI systems through XAI influences how individuals evaluate and respond to AI-assisted hiring recommendations. Specifically, the study investigated the effects of XAI, compared to blackbox AI, on perceived explainability, trust, perceived fairness, and willingness to follow AI recommendations, as well as the mediating role of explainability and trust in shaping behavioral acceptance.

Overall, the findings provide strong and consistent evidence that explainability matters for how AI systems are perceived in recruitment contexts. In line with H1 and H2, participants exposed to the XAI system reported significantly higher levels of trust and perceived explainability compared to those exposed to a blackbox AI. The effect of XAI on perceived explainability was particularly strong, indicating that participants exposed to the transparent system reported substantially higher levels of explainability than those exposed to the opaque system and recognized the additional information provided by the explainable model.

Regarding fairness perceptions, the results also support H3. Participants in the XAI condition perceived the AI-assisted hiring process as significantly fairer than participants in the blackbox condition, even after controlling for demographic characteristics and prior experience with AI tools. This suggests that providing explanations not only improves understanding but also enhances perceptions of procedural justice in algorithmic decision-making. At the same time, individual differences played a role: male participants reported higher fairness perceptions, while participants with greater experience with AI tended to evaluate the system more critically.

In contrast, the direct relationship between AI system type and behavioral acceptance of the AI recommendation was weaker. Consistent across both bivariate correlations and regression-based analyses, exposure to XAI did not lead to a statistically significant increase in willingness to follow the AI's recommendation, providing no direct support for H4. Nevertheless, the total effect of AI system type on recommendation acceptance was positive and marginally significant ($p = .080$), suggesting a tentative tendency toward greater acceptance in the XAI condition. Overall, this pattern This indicates that transparency alone is not sufficient to immediately change users' decision behavior in a sensitive context such as recruitment.

Crucially, the mediation analysis provided important insight into this pattern. Results partially supported H5, showing that the effect of AI system type on recommendation acceptance operated indirectly through trust, but not through perceived explainability. While XAI substantially increased explainability, only trust emerged as a significant mediator linking AI type to behavioral acceptance. In other words, understanding how the AI works did not, by itself, lead participants to rely more on the recommendation; instead, it was the trust inspired by the system that translated transparency into behavioral change.

Taken together, these findings point to an important distinction between how people evaluate AI systems and how they act upon their recommendations. XAI improves users' perceptions of

transparency, trust, and fairness, but behavioral acceptance appears to depend primarily on whether the system is trusted. This highlights the central role of trust as the bridge between transparency and action in AI-assisted recruitment decisions.

5.2 Implications

The findings of this dissertation offer several important theoretical and practical implications for research on XAI and its application in recruitment contexts.

From a theoretical perspective, this study contributes to the growing literature on human-AI interaction by empirically disentangling the roles of perceived explainability and trust in shaping users' responses to AI-assisted hiring decisions. While prior research has frequently conceptualized explainability, trust, and acceptance as closely intertwined or mutually reinforcing constructs (e.g., Weitz et al., 2019; Dwivedi et al., 2023), the present findings demonstrate that these constructs play distinct and non-substitutable roles. Specifically, although exposure to XAI substantially increased perceived explainability and trust, only trust emerged as a significant mediator of behavioral acceptance. This finding refines existing theoretical models by showing that explainability alone does not automatically translate into action; rather, trust serves as the key psychological mechanism through which transparency becomes behaviorally meaningful. In doing so, the study supports and extends trust-based frameworks of automation reliance (Rosenbacke et al., 2024), highlighting trust as the critical bridge between system characteristics and user behavior in high-stakes decision-making contexts such as recruitment.

Moreover, the strong effect of XAI on perceived fairness provides empirical support for procedural justice perspectives in algorithmic decision-making discussed in the literature review. Prior research has argued that limited transparency in algorithmic systems undermines perceptions of fairness, ethics, and legitimacy in AI-assisted hiring (Chamorro-Premuzic et al., 2023), whereas greater insight into decision processes supports fairness judgments (Hofeditz et al., 2022; Ramachandranpillai et al., 2023). By demonstrating that explainable systems are perceived as significantly fairer than blackbox systems - even after controlling for demographic variables and prior AI experience - this study reinforces the theoretical claim that transparency enhances perceived procedural justice and legitimacy in automated recruitment. Importantly, the findings extend prior work by showing that fairness perceptions can be meaningfully

improved through relatively simple explanatory interventions, without modifying the underlying decision outcome itself (Deck et al., 2024).

From a practical perspective, the results offer valuable guidance for organizations implementing AI in recruitment. First, the findings suggest that incorporating explanations into AI-assisted hiring tools can meaningfully improve users' trust and fairness perceptions. This aligns with prior work emphasizing the importance of human-readable explanations for reducing uncertainty and fostering appropriate trust in AI systems (De Brito Duarte et al., 2023). Notably, even brief explanations of decision criteria were sufficient to make the system feel more transparent and legitimate, which is particularly relevant in recruitment contexts where decisions are sensitive and have substantial personal and organizational consequences.

At the same time, the absence of a direct effect of XAI on willingness to follow AI recommendations underscores an important managerial insight already anticipated in the literature. As discussed by Fleiß et al. (2024) and Kalff & Simbeck (2025), acceptance of AI advice is highly context-dependent and cannot be assumed to follow automatically from system transparency. Importantly, in the present study, the AI system consistently provided recommendations aligned with participants' strongest profile, thereby representing high-quality advice. Future research should therefore examine contexts in which AI recommendations are biased, incorrect, or misaligned with user interests in order to assess whether explainability effectively supports calibrated trust and appropriate reliance. The present findings reinforce the view that explainability should not be treated as a mechanism to compel compliance with AI recommendations. Instead, its primary value lies in supporting informed human judgment and fostering calibrated trust, rather than replacing human decision-making authority. Organizations should therefore position XAI as a decision-support tool that complements human expertise, rather than as an authoritative system whose outputs must be followed uncritically.

Finally, the finding that trust - but not perceived explainability - mediated behavioral acceptance has important implications for the design and governance of AI systems in organizations. Consistent with arguments by Hamm et al. (2023) and Bernardo & Seva (2023), the results suggest that transparency constitutes a necessary but insufficient condition for AI reliance. Organizations seeking to increase meaningful use of AI systems should therefore focus not only on making systems understandable, but also on cultivating broader trust in their reliability,

competence, and alignment with organizational values. Transparency lays the foundation for trust, but trust is the mechanism through which transparency translates into behavior.

5.3 Limitations

Like any research project, this dissertation comes with several limitations that should be acknowledged. The first relates to the experimental nature of the study. Although the task was designed to closely resemble realistic recruitment scenarios and thus offered a high degree of procedural realism, participants evaluated fictional applicants in a controlled, consequence-free environment. As a result, while decision processes may approximate those used in real hiring contexts, the study cannot fully capture the pressures, accountability, and ethical weight associated with actual personnel decisions. In real-world settings, where decisions have tangible consequences for candidates and organizations, individuals may engage more cautiously or strategically with AI-supported recommendations.

Another limitation concerns the sample characteristics. Most of the participants were young, highly educated, and based in Portugal. While this group offers useful insights, it does not reflect the diverse perspectives of experienced hiring managers, HR professionals, or candidates from different cultural backgrounds. Perspectives on AI—especially concerning fairness and trust—can vary greatly depending on a person's professional experience, cultural background, or degree of technological understanding (Chamorro-Premuzic et al., 2023; Hofeditz et al., 2022; Kalff & Simbeck, 2025).

A third limitation relates to the measurement of behavioral acceptance. In this study, acceptance was captured through the change in participants' choices before and after seeing the AI recommendation. Although this measure helps quantify shifts in decision-making, it cannot fully represent the nuanced way people often negotiate between their own judgment and algorithmic advice. Prior research suggests that acceptance of AI recommendations often involves deliberation, selective reliance, and contextual trade-offs that may not be fully captured by simple directional change measures (Hamm et al., 2023; Bernardo & Seva, 2023; Fleiß et al., 2024).

The study also faced a measurement limitation in the fairness scale, which showed a Cronbach's α below .70 (Bland & Altman, 1997). While the scale still provided useful information, its reliability was lower than ideal. This suggests that participants' views of fairness might not have been recorded with the accuracy or consistency sought. Future studies ought to enhance

fairness metrics or develop frameworks tailored specifically for AI-based hiring contexts, where fairness represents a particularly delicate and multidimensional concept.

Finally, the study used only one type of explanation in the XAI condition—a short textual rationale. Although this choice allowed for experimental control, different explanation styles may influence users in different ways. Prior research has shown that explanation format, level of detail, and presentation style can shape users' perceptions of trust, fairness, and usefulness in distinct ways. By focusing on a single explanation format, the study cannot determine whether alternative forms of explainability might have produced stronger—or weaker—effects (Weitz et al., 2019; De Brito Duarte et al., 2023; Rosenbacke et al., 2024).

5.4 Directions for Future Research

Given these limitations, various potential directions for future research can be identified. An important factor is the need to study XAI in real organizational contexts; particularly, witnessing the dynamics between hiring managers or candidates and explainable AI in authentic recruitment processes would provide a more profound understanding of how transparency affects real actions instead of merely theoretical choices.

Future research would also gain from having more varied samples. Including professional recruiters, job seekers in real selection processes, or participants from different countries would deepen our knowledge about how the concepts of trust, fairness, and explainability are viewed across diverse groups. Such diversity might uncover cultural or professional patterns that the current study could not capture.

Another productive direction would be to create more granular behavioral measures: Rather than simply measuring preference shifts, for example, future work might collect decision justifications, think-aloud data, or digital traces that offer a window into how people interpret and integrate algorithmic recommendations. These techniques may clarify the mental processes that underlie agreement or opposition to AI recommendations.

Due to the reliability concerns associated with the fairness scale in this study, future research should focus on developing more robust and contextually relevant fairness metrics. Equity is an essential issue in algorithmic hiring, yet it remains a complex and multi-faceted notion. Differentiating between procedural fairness, distributive fairness, and interpersonal fairness could offer a better understanding of how individuals assess decisions made by AI.

Future work should also consider other explanation formats within XAI. Counterfactuals, visual models, feature-importance analyses, and hybrid explanations might engage users in distinct ways, influencing explainability, trust, and fairness in various manners. Grasping which explanations are most effective - and for which individuals - is essential for creating successful XAI systems.

Collectively, these guidelines highlight the necessity to persist in exploring not just if XAI enhances human-AI interaction, but also how particular psychological and contextual elements influence the responses of individuals to algorithmic suggestions.

5.5 Conclusion

The focus of this research has been to explore how users' perceptions and intentions to use AI-supported recruitment decisions can be improved by either an XAI model as compared to the blackbox model of AI. Through the experimental framework of this research, the concerns that have been emphasized within the existing body of knowledge regarding the perceived significance of clarity within AI models have been taken up by this research to address how recruitment-related decisions within AI can actually be perceived and judged as has been proposed within the existing body of knowledge by Weitz et al. (2019), as well as Dwivedi et al. (2023).

The results indicate that XAI had a significant positive impact on perceived explainability, trust, and perceived fairness, corroborating the assumption in the earlier works stating that understanding the workings of an algorithmic system diminishes ambiguity and facilitates well-educated trust decisions (De Brito Duarte et al., 2023; Rosenbacke et al., 2024). Yet, the results indicate that the mere factor of transparency played no role in influencing users to act on the advice of the AI system directly. Trust acts as the mediating factor in the explainability-behavioral acceptance relationship, developing the conceptualization of understanding an AI system and the reliance on the system within decision-making processes, aligning well with the claims that explainability and acceptance are not synonymous outcomes (Hamm et al., 2023; Bernardo & Seva, 2023).

Overall, these findings indicate that XAI has a critical role in enhancing transparency and trust in AI-assisted recruitment practices in line with procedural justice visions that associate transparency and trust in AI decision-making systems (Chamorro-Premuzic et al., 2023; Hofeditz et al., 2022). Nonetheless, these findings are also consistent with visions regarding the

basic role of XAI as a means that assists human decision-making rather than automating it (Fleiß et al., 2024; Kalff & Simbeck, 2025). In this context, given that AI applications are now becoming ubiquitous in human resource management practices in organizations, these findings highlight the need for trust-building or transparency-enhancing in AI-related systems in this domain if these systems are going to be used effectively in a long-term manner.

6. References

- Abbass, H. A. (2019). “Social Integration of Artificial Intelligence: Functions, Automation Allocation Logic and Human-Autonomy Trust”. English. In: *Cognitive Computation*, 10(48), pp. 1–13.
- Ahmed, W. (2024). Feature Creation to Enhance Explainability and Predictability of ML Models Using XAI. *International Journal of Advanced Computer Science & Applications*, 15(10).
- Amitabh, U., & Ansari, A. (2025, March 28). Hiring with AI doesn’t have to be so inhumane. *World Economic Forum*. https://www.weforum.org/stories/2025/03/ai-hiring-human-touch-recruitment/
- Angelopoulos, G., Lacroix, D., Wullenkord, R., Rossi, A., Rossi, S., & Eyssel, F. (2025). Measuring transparency in intelligent robots. *Scientific Reports*.
- Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5), e1424.
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges towardresponsible AI. *Information Fusion*, 58, 82-115. https://doi.org/10.1016/j.inffus.
- Baert, S. (2018). Hiring discrimination: An overview of (almost) all correspondence experiments since 2005. Springer, Cham.
- Baert, S., Albanese, A., du Gardein, S., Ovaere, J., & Stappers, J. (2017). Does work experience mitigate discrimination?. *Economics Letters*, 155, 35-38.
- Baert, Stijn (2017) : Hiring Discrimination: An Overview of (Almost) All Correspondence Experiments Since 2005, *IZA Discussion Papers*, No. 10738, Institute of Labor Economics (IZA), Bonn.
- Band, S. S., Yarahmadi, A., Hsu, C. C., Biyari, M., Sookhak, M., Ameri, R., ... & Liang, H. W. (2023). Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked*, 40, 101286.
- Bernardo, E. L., & Seva, R. R. (2023, December). Exploration of explainable AI for trust development on human-AI interaction. In *Proceedings of the 2023 6th Artificial Intelligence and Cloud Computing Conference* (pp. 238-246).

- Bertrand, A., Viard, T., Belloum, R., Eagan, J. R., & Maxwell, W. (2023, April). On selective, mutable and dialogic xai: A review of what users say about different types of interactive explanations. In *Proceedings of the 2023 CHI conference on human factors in computing systems* (pp. 1-21).
- Black, J. S., & van Esch, P. (2020). ai-enabled recruiting: What is it and how should a manager use it? *Business Horizons*, 63(2), 215–226. https://doi.org/10.1016/j.bushor.2019.12.001
- Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *Bmj*, 314(7080), 572.
- Bringsjord, S. & Schimanski, B. (2003). What is Artificial Intelligence? Psychometric AI as an answer. *IJCAI'03 Proceedings of the 18th international joint conference on Artificial intelligence*, 887-893
- Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*, 109, 104047. https://doi.org/10.1016/j.jml.2019.104047
- Bucher, T. (2016). Neither black nor box: Ways of knowing algorithms. In *Innovative methods in media and communication research*, 81-98. Palgrave Macmillan, Cham.
- Cambridge University Press & Assessment. (2025). Artificial. In *Cambridge Dictionary*. Retrieved December 30, 2025, from https://dictionary.cambridge.org/dictionary/english/artificial
- Capone, L., & Bertolaso, M. (2020). A philosophical approach for a human-centered explainable AI. In In *Proceedings of the Italian Workshop on Explainable Artificial Intelligence co-located with 19th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2020)* (Vol. 2742, pp. 80-86).
- Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of economic behavior & organization*, 81(1), 1-8.
- Chamorro-Premuzic, T., Akhtar, R., Winsborough, D., & Sherman, R. A. (2023). The impact of AI on recruitment fairness and transparency. *Journal of Applied Psychology*, 108(1), 11–27.
- Cheung, J. C., & Ho, S. S. (2025). The effectiveness of explainable AI on human factors in trust models. *Scientific Reports*, 15(1), 23337.

- Colquitt, J. A., & Rodell, J. B. (2015). Measuring justice and fairness. *The Oxford handbook of justice in the workplace*, 1, 187-202.
- Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*(Vol. 1195). Boston, MA: Houghton Mifflin.
- Davenport, T. H. (2018b). *The AI Advantage: How to Put the Artificial Intelligence Revolution to Work*. Cambridge: MIT Press.
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7), 571–582. https://doi.org/10.1037/0003-066X.34.7.571
- Deck, L., Schoeffer, J., De-Arteaga, M., & Kühl, N. (2024, June). A critical survey on fairness benefits of explainable AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1579-1595).
- DeSimone, H., & Leon-Espinosa, M. (2024, March). Explainable ai: The quest for transparency in business and beyond. In *2024 7th International Conference on Information and Computer Technologies (ICICT)* (pp. 532-538). IEEE.
- de Brito Duarte, R., Correia, F., Arriaga, P., & Paiva, A. (2023). AI trust: Can explainable AI enhance warranted trust?. *Human Behavior and Emerging Technologies*, 2023(1), 4637678.
- Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., ... & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM computing surveys*, 55(9), 1-33.
- Fleiß, J., Bäck, E., & Thalmann, S. (2024). Mitigating algorithm aversion in recruiting: A study on explainable AI for conversational agents. *Acm sigmis database: the database for advances in information systems*, 55(1), 56-87.
- Fuller, J., Raman, M., Sage-Gavin, E., & Hines, K. (2021, September). *Hidden Workers: Untapped Talent*. Harvard Business School Project on Managing the Future of Work & Accenture. https://www.hbs.edu/managing-the-future-of-work/Documents/research/hiddenworkers09032021.pdf
- Ghoneim, M. (2024). Strategic integration of Explainable AI in Recruitment: Towards a framework for Ethical decision-making and Stakeholder confidence.
- Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, October). Explaining explanations: An overview of interpretability of machine

learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)* (pp. 80-89). IEEE.

- Gino, F., & Moore, D. (2007). Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making*, 20, 21–35.
- Gopi Krishna, D., Ajay Kanth, K., Kiruthika, M., & Sheryl Oliver, A. (2025). A theoretical framework for visual representation and user interface design in explainable AI for HR management using machine learning models. In *Proceedings of the 6th International Conference on Information Management & Machine Intelligence (ICIMMI 2024)*(pp. 1–8). https://doi.org/10.1145/3745812.3745837
- Haefner, N., Wincent, J., Parida, V. & Gassmann, O. (2021). Artificial intelligence and innovation management: A review, framework, and research agenda. *Technological Forecasting & Social Change*, 162. 120392. https://doi.org/10.1016/j.techfore.2020.120392
- Haenlein, M. & Kaplan, A. (2019). A Brief History of Artificial Intelligence: On the Past, Present and Future of Artificial Intelligence. *California Management Review* 61(4): pp. 5 -14. https://doi.org/10.1177/0008125619864925
- Hamm, P., Klesel, M., Coberger, P., & Wittmann, H. F. (2023). Explanation matters: An experimental study on explainable AI. *Electronic Markets*, 33(1), 17.
- Hamida, S. U., Chowdhury, M. J. M., Chakraborty, N. R., Biswas, K., & Sami, S. K. (2024). Exploring the Landscape of Explainable Artificial Intelligence (XAI): A Systematic Review of Techniques and Applications. *Big Data and Cognitive Computing*, 8(11), 149. https://doi.org/10.3390/bdcc8110149
- Hauptman, A. I., Schelble, B. G., Duan, W., Flathmann, C., & McNeese, N. J. (2024). Understanding the influence of AI autonomy on AI explainability levels in human-AI teams using a mixed methods approach. *Cognition, Technology & Work*, 26(3), 435-455.
- Hausknecht, J., Day, D., & Thomas, S. 2004. Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology*, 57: 639-684.
- Hofeditz, L., Clausen, S., Rieß, A., Mirbabaie, M., & Stieglitz, S. (2022). Applying XAI to an AI-based system for candidate management to mitigate bias and discrimination in hiring. *Electronic Markets*, 32(4), 2207-2233.

- Hoff, K. A., and Bashir, M. (2015). Trust in automation: integrating empirical evidence on factors that ifnuence trust. *Hum. Factors* 57, 407–434. doi: 10.1177/0018720814547570
- Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4), e1312.
- Hurlburt, G. (2017) 'How much to trust Artificial Intelligence?', *IT Professional*, 19(4), pp. 7–11, EEE Explore. doi:10.1109/mitp.2017.3051326
- Iansiti, M., & Lakhani, K. R. (2020). *Competing in the age of AI: Strategy and leadership when algorithms and networks run the world*. Harvard Business Press.
- Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine Learning and Deep Learning. *Electronic Markets*, 31(3), 685–695. https://doi.org/10.1007/s12525-021-00475-2
- Jatobá, M., Santos, J., Gutierriz, I., Moscon, D., Fernandes, P. O., & Teixeira, J. P. (2019). Evolution of artificial intelligence research in human resources. *Procedia Computer Science*, 164, 137-142.
- John-Mathews, J. M. (2021). Critical empirical study on black-box explanations in AI. *arXiv preprint*arXiv:2109.15067.
- Johnson, J. (2020). Interpretability vs explainability: The black box of machine learning. *BMC Blogs*.
- Kalff, Y., & Simbeck, K. (2025). Explained, yet misunderstood: How AI Literacy shapes HR Managers' interpretation of User Interfaces in Recruiting Recommender Systems. *arXiv preprint* arXiv:2509.06475.
- Kaplan, J. (2016). *Artificial Intelligence: What everyone needs to know*. Oxford: Oxford university press
- Kemppi, J. 2018. Tekoälyn mahdollisuudet ovat rajattomat, mutta myös uhkialöytyy: ”Jos asiassa olla tarkkana, voi koitua suuria ongelmia”. *Iltalehti*. https://www.iltalehti.fi/digiuutiset/a/201805042200921731
- Kolbjørnsrud, V. (2017). Kunstig Intelligens og Lederens Nye Jobb. *Magma*, 20(06), 33-42.
- Kolbjørnsrud, V., Sannes, R. (2022). Problemløsing med kunstig intelligens: Bruk av spacemaker i tidligfase eiendomsutvikling. *Praktisk økonomi & finans*, 38(1). 47-64. https://doi.org/10.18261/pof.38.1.4

- Lacroux, A., & Martin-Lacroux, C. (2022). Should I trust the artificial intelligence to recruit? Recruiters' perceptions and behavior when faced with algorithm-based recommendation systems during resume screening. *Frontiers in Psychology*, 13, 895997.
- Langer, M., & König, C. (2022). Explainability of artificial intelligence in human resources. In *Handbook of research on artificial intelligence in human resource management* (pp. 285-302). Edward Elgar Publishing.
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1), 50-80.
- Legg, S. & Hutter, M. (2007). A Collection of Definitions of Intelligence. *Advances in Artificial General Intelligence: Concepts, Architecture and Algorithms*, IOS press, Issue 157, 17-24
- Lin, Z. Q., Shafiee, M. J., Bochkarev, S., Jules, M. S., Wang, X. Y., & Wong, A. (2019). Do explanations reflect decisions? A machine-centric strategy to quantify the performance of explainability algorithms. *arXiv preprint*arXiv:1910.07387.
- Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1), 18.
- LinkedIn Engineering. (2021). Transparent and Explainable AI Systems. LinkedIn blog https://www.linkedin.com/blog/engineering/fairness/transparent-and-explainable-ai-systems
- Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3), 31-57.
- Lundvall, H. (2022). Artificial Intelligence in Recruitment: Opportunities and Challenges of Implementing Artificial Intelligence in today's Recruitment Processes.
- Marey, A., Arjmand, P., Alerab, A. D. S., Eslami, M. J., Saad, A. M., Sanchez, N., & Umair, M. (2024). Explainability, transparency and black box challenges of AI in radiology: impact on patient care in cardiovascular radiology. *Egyptian Journal of Radiology and Nuclear Medicine*, 55(1), 183.
- Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of biomedical informatics*, 113, 103655.

- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, 20(3), 709-734.
- Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press. https://doi.org/10.1037/11281-000
- Merritt, S. M. (2011). Affective Processes in Human-Automation Interactions. *Human Factors*, 53, 356- 370.
- Mirbabaie, M., Brünker, F., Möllmann Frick, N. R. J., & Stieglitz, S. (2021). The Rise of Artificial Intelligence – Understanding the AI Identity Threat at the Workplace. *Electronic Markets*, 32, 73–99. https://doi.org/10.1007/s12525-021-00496-x
- Nakao, Y., Stumpf, S., Ahmed, S., Naseer, A., & Strappelli, L. (2022). Toward involving end-users in interactive human-in-the-loop AI fairness. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 12(3), 1-30.
- Nechytailo, A. (2023). Using AI-powered tools for Improving Recruitment Processes.
- Nguyen, Q. T., Le, L., Tran, X. T., Do, T., & Lin, C. T. (2024, October). FairAD-XAI: Evaluation Framework for Explainable AI Methods in Alzheimer's Disease Detection with Fairness-in-the-loop. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 870-876).
- Nicolás-Agustín, Á., Jiménez-Jiménez, D., & Maeso-Fernandez, F. (2022). The role of human resource practices in the implementation of digital transformation. *International Journal of Manpower*, *43*(2), 395-410.
- Ochmann, J., & Laumer, S. (2019). Fairness as a determinant of AI adoption in recruiting: An interview-based study.
- Ore, O., & Sposato, M. (2022). Opportunities and risks of artificial intelligence in recruitment and selection. *International Journal of Organizational Analysis*, 30(6), 1771-1782. https://doi.org/10.1108/IJOA-07-2020-2291
- Ortigossa, E. S., Gonçalves, T., & Nonato, L. G. (2024). Explainable artificial intelligence (xai)—from theory to methods and applications. *IEEE Access*, 12, 80799-80846.
- Papanikou, V., Karidi, D. P., Pitoura, E., Panagiotou, E., & Ntoutsi, E. (2025). Explanations as Bias Detectors: A Critical Study of Local Post-hoc XAI Methods for Fairness Exploration. *arXiv preprint* arXiv:2505.00802.

- Parry, K., Cohen, M. & Bhattacharya, S. (2016). Rise of the Machines: A Critical Consideration of Automated Leadership Decision Making in Organizations. *Group & Organization Management*, 41(5). 571-594. https://doi.org/10.1177/1059601116643442
- Petrat, D. (2021). Attitude Towards Artificial Intelligence in a Leadership Role. In Black, N.L., Neumann, W.P., Noy, I. (eds.), *Proceedings of the 21st Congress of the International*
- Quinn, T. P., Jacobs, S., Senadeera, M., Le, V., & Coghlan, S. (2022). The three ghosts of medical AI: Can the black-box present deliver? *Artificial intelligence in medicine*, 124, 102158.
- Rankila, E. (2020). Artificial intelligence (AI) in recruitment.
- Ramachandranpillai, R., Baeza-Yates, R., & Heintz, F. (2023). FairXAI-A Taxonomy and Framework for Fairness and Explainability Synergy in Machine Learning. *IEEE Transactions on Neural Networks and Learning Systems*.
- Rasmussen, J., Pejterson, A. M., & Goodstein, L. P. (1994). *Cognitive systems engineering*. New York: Wiley.
- Rastogi, N., Dhanuka, D., Saxena, A., Mairal, P., & Nguyen, L. (2025). Survey Perspective: The Role of Explainable AI in Threat Intelligence. *arXiv preprint* arXiv:2503.02065.
- Rigotti, C., & Fosch-Villaronga, E. (2024). Fairness, AI & recruitment. *Computer Law & Security Review*, 53, 105966. https://doi.org/10.1016/j.clsr.2024.105966
- Rosenbacke, R., Melhus, Å., McKee, M., & Stuckler, D. (2024). How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: systematic review. *Jmir Ai*, 3, e53207.
- Salin, E. D., & Winston, P. H. (1992). Machine Learning and Artificial Intelligence. *Analytical chemistry*, 64(1).
- Schemerhorn J.R (2001). *Management Updated*. New York: John Wiley and Sons. Inc.
- Schweizer, K. (2011). Some thoughts concerning the recent shift from measures with many items to measures with few items. *European Journal of Psychological Assessment*, 2Z, 71-72. doi: 10.1027/1015-5759/a000056

- Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, 146, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551
- Shulner-Tal, A., & Sheidin, J. (2025, June). XAI4RE–Using Explainable AI for Responsible and Ethical AI. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization* (pp. 300-307).
- Sim, M., Kim, S., & Suh, Y. (2022). Sample size requirements for simple and complex mediation models. *Educational and Psychological Measurement*, 82(1), 76–106. https://doi.org/10.1177/001316442110032.
- Singh, P., & Finn, D. (2022). Understanding applicant perspectives of AI-based recruitment: A review and research agenda. *Journal of Business and Psychology*, 37(4), 833–848.
- Singh, R., & Finn, A. (2022). Artificial intelligence in recruitment: Examining candidate perceptions and organizational outcomes. *Human Resource Management Review*, 32(3), 100877.
- Storey, D.J. (2004). Exploring the link, among small firms, between management training and firm performance: a comparison between the UK and other OECD countries. *The international Journal of Human Resource Management*, 15(1)
- Stuart, R., & Norvig, P. (2016). *Artificial Intelligence: A Modern Approach*. Third Edition. Prentice Hall Press Upper Saddle River.
- Tabassam, A., Yaqoob, G., Cuong, V. H., Syed, M., Shahzadi, A., & Asghar, F. (2023). The Ethical Implication of Using Artificial Intelligence in Hiring and Promotion Decisions. *Journal of Management & Educational Research Innovation*, 1(2), 1-15. https://doi.org/10.5281/zenodo.10066900
- Tambe, P., Cappelli, P. & Yakubovich, V. (2019). Artificial Intelligence in Human Resources Management: Challenges and a Path Forward. *California Management Review*, 61(4), 15-42
- Tecuci, G. (2012). Artificial Intelligence. *Wires computational statistics*, 4(2), 168-180
- Thakur, Y., Dash, A., Patel, A. K., & Sayyad, A. (2025). A Comprehensive Review of Explainable AI (XAI) Methods in Deep Learning. *International Journal of Scientific Research in Science and Technology*, 12(5), 333-352.

- Titareva, T. (2021). Leadership in an Artificial Intelligence Era. Paper for Leading Change Conference 2021. School of Strategic Leadership Studies, James Madison University.
- van Esch, Patrick, J. Stewart Black, and Joseph Ferolie. 2019. Marketing AI recruitment: the next phase in job application and selection. *Computers in Human Behavior* 90: 215–222.
- Ved, S., Kaundanya, N.S. & Panda, O.P. (2016). Applications and Current Achievements in the field of Artificial Intelligence. *Imperial Journal of Interdisciplinary research*, 2(11), 932-936
- Vultureanu-Albişi, A., & Bădică, C. (2022). A survey on effects of adding explanations to recommender systems. *Concurrency and Computation: Practice and Experience*, 34(20), e6834. https://doi.org/10.1002/cpe.6834
- Wall T.D. & Wood S.J. (2005). The romance of human resource management and business performance, and the case for big science. *Human Relations*, 58(4), 429-462
- Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019, May). Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1-15).
- Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2019, July). " Do you trust me?" Increasing user-trust by integrating virtual agents in explainable AI interaction design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (pp. 7-9).
- Wesche, J. S., Hennig, F., Kollhed, C. S., Quade, J., Kluge, S., & Sonderegger, A. (2024). People's reactions to decisions by human vs. algorithmic decision-makers: The role of explanations and type of selection tests. *European Journal of Work and Organizational Psychology*, 33(2), 146-157.
- Wijayati, D. T., Rahman, Z., Fahrullah, A., Rahman, M. F. W., Arifah, I. D. C & Kautsar, A. (2022). A Study of Artificial Intelligence on Employee Performance and Work Engagement: The Moderating Role of Change Leadership. *International Journal of Manpower*, 43(2), 486- 512. https://doi.org/10.1108/IJM-07-2021-0423
- Williams, O. (2021). Towards human-centred explainable AI: A systematic literature review. Master's Thesis.
- Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational behavior and human decision processes*, 93(1), 1-13.

- Zhang, G., Pan, L., Tang, F., & Yao, F. (2025). Explainable artificial intelligence in the talent recruitment process-a literature review. *Cogent Business & Management*, 12

7. Appendix

Appendix A: *Qualtrics Survey*

English

Introduction and Consent

Welcome and thank you for considering taking part in this study. My name is Filipa Duarte, and I am conducting this experiment as part of my Master's Thesis at the Católica Lisbon School of Business and Economics, under the supervision of Professor Cristina Mendonça.

This study involves reading a short scenario that includes an Artificial Intelligence (AI) system and then answering some questions about your impressions and decisions. It should take approximately 6-8 minutes to complete. Please answer as honestly as possible. We kindly ask that you complete the study in one sitting, without interruptions.

The purpose of this research is to better understand how people interact with AI in recruitment decision contexts. Your participation will help advance research

on human-AI interaction and decision-making.

There are no expected side effects of participating in this study beyond those associated with looking at a computer screen for some minutes. All responses will be treated with strict confidentiality and are completely anonymous. This means that it will not be possible to link your answers to your identity in any way. The data will be used for research purposes only and may appear in my thesis or academic publications, always in aggregated form, never in a way that could identify individual participants.

You may change your mind and drop out at any point of the study during its completion.

If you have any questions or concerns about this research, please contact me at s-fialduarte@ucp.pt.

By clicking "Continue," you acknowledge that you have read this information and agree to participate in the study.

Thank you very much for your time and participation!

Company name

Please enter the name of a company you currently work for, or a company you would like to work for in the future.

Recruitment Decision Scenario

Imagine that you are an HR manager in ${q://QID7/ChoiceTextEntryValue} evaluating two candidates for a Data Scientist position.

The profiles include typical résumé information (education, experience, technical/soft skills, and prior performance). Please review both candidates carefully.

You will be asked for your preference before and after seeing an AI recommendation.

Initial Preference (before AI recommendation)

Please review the two candidate profiles below.
Sam Jones

- **Education:** Master's degree in Data Science
- **Experience:** 4 years of experience in data analytics

and machine learning
- **Technical skills:** Python, R, PowerBI, TensorFlow
- **Projects:** Developed a predictive sales model that improved forecast accuracy by 15%
- **Soft skills:** Strong analytical thinking, communication, and problem-solving abilities

Jess Smith

- **Education:** Bachelor's degree in Statistics
- **Experience:** 1 year of experience in data analytics
- **Technical skills:** Excel, PowerBI
- **Projects:** Assisted in preparing monthly sales reports
- **Soft skills:** Team-oriented, motivated to learn new tools

Please review the two candidate profiles below.

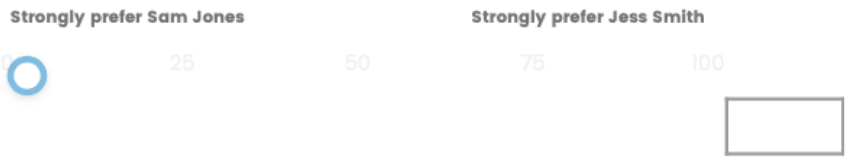
Sam Jones

- **Education:** Bachelor's degree in Statistics
- **Experience:** 1 year of experience in data analytics
- **Technical skills:** Excel, PowerBI Projects: Assisted in preparing monthly sales reports
- **Soft skills:** Team-oriented, motivated to learn new tools

Jess Smith

- **Education:** Master's degree in Data Science
- **Experience:** 4 years of experience in data analytics and machine learning
- **Technical skills:** Python, R, PowerBI, TensorFlow
- **Projects:** Developed a predictive sales model that improved forecast accuracy by 15%
- **Soft skills:** Strong analytical thinking, communication, and problem-solving abilities

Based on the information above, which candidate would you prefer to hire? Assuming that 0=Strongly prefer Sam Jones and 100=Strongly prefer Jess Smith.



If blackbox + Sam strong/Jess weak

Please imagine the company ${q://QID7/ChoiceTextEntryValue} uses an AI decision-support system that analyzes candidate profiles and

provides a recommendation score on a 0-100 scale, where, in this case, 0 = Strongly prefers Sam Jones and 100 = Strongly prefers Jess Smith.

For these two candidates, the output of the AI system was:
AI Screening Score: 25
AI Recommendation: Sam Jones

If XAI + Sam strong/Jess weak

Please imagine the company ${q://QID7/ChoiceTextEntryValue} uses an AI decision-support system that analyzes candidate profiles and provides a recommendation score on a 0-100 scale, where, in this case, 0 = Strongly prefers Sam Jones and 100 = Strongly prefers Jess Smith.

For these two candidates, the output of the AI system was:
AI Recommendation Score: 25
AI Recommendation: Sam Jones
AI Explanation: Sam Jones scored higher due to their graduate education in Data Science, four years of relevant experience, and advanced technical skills (Python, Power BI, TensorFlow) that align with the job

requirements. Jess Smith scored lower because their education and experience are more limited, and their technical background is less specialized for this role.

If blackbox + Sam weak/Jess strong

Please imagine the company ${q://QID7/ChoiceTextEntryValue} uses an AI decision-support system that analyzes candidate profiles and provides a recommendation score on a 0-100 scale, where, in this case, 0 = Strongly prefers Sam Jones and 100 = Strongly prefers Jess Smith.

For these two candidates, the output of the AI system was:
AI Screening Score: 75
AI Recommendation: Jess Smith

If XAI + Sam weak/Jess strong

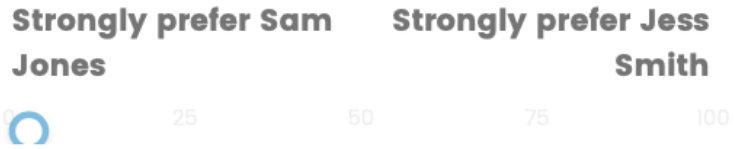
Please imagine the company ${q://QID7/ChoiceTextEntryValue} uses an AI decision-

support system that analyzes candidate profiles and provides a recommendation score on a 0-100 scale, where, in this case, 0 = Strongly prefers Sam Jones and 100 = Strongly prefers Jess Smith.

For these two candidates, the output of the AI system was:
AI Screening Score: 75
AI Recommendation: Jess Smith
AI Explanation: Jess Smith scored higher due to their graduate education in Data Science, four years of relevant experience, and advanced technical skills (Python, Power BI, TensorFlow) that align with the job requirements. Sam Jones scored lower because their education and experience are more limited, and their technical background is less specialized for this role.

Final preference (After AI recommendation)

Considering the AI system's recommendation, which candidate would you now prefer to hire? Assuming that 0=Strongly prefer Sam Jones and 100 = Strongly prefer Jess Smith.



Scales

Please rate how much you agree with the following statements about the AI systems in recruitment and selection.

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
I usually trust AI systems in recruitment decisions until there is a reason not to.	○	○	○	○	○
For the most part, I distrust AI systems used in hiring decisions.	○	○	○	○	○
In general, I would rely on an AI system to assist in evaluating job candidates.	○	○	○	○	○
My tendency to trust AI systems in recruitment is high.	○	○	○	○	○
It is easy for me to trust an AI system to do its job in candidate selection.	○	○	○	○	○
I am likely to trust an AI system even when I have little knowledge about it.	○	○	○	○	○

Please rate how clear and understandable you found the AI system's recommendation.

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
The AI system's explanations are useful.	○	○	○	○	○
The AI system provides clear explanations for its recommendations.	○	○	○	○	○
I feel informed about the AI system's decisions.	○	○	○	○	○
The AI system conveys its reasoning effectively.	○	○	○	○	○

Please rate how fair and appropriate you found the AI system's evaluation.

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
The AI system used for candidate selection is free of bias.	○	○	○	○	○
The AI system upholds ethical and moral standards.	○	○	○	○	○
It is fair that the AI considers candidate characteristics (e.g., experience, education, skills) when evaluating them.	○	○	○	○	○

Please rate how effective and useful you think AI systems are in recruitment.

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strong disagr
I can see the benefits of AI systems that can make hiring decisions better than humans.	○	○	○	○	○
AI systems that can assist in recruitment could be useful.	○	○	○	○	○
I believe this kind of AI system can perform well in evaluating job candidates.	○	○	○	○	○

Please indicate how you feel about the use of AI in recruitment.

	Strongly agree	Somewhat agree	Neitgher agree nor disagree	Somewhat disagree	Stro disc
AI systems that can make hiring decisions better than humans make me uncomfortable.	○	○	○	○	(
AI systems that can evaluate people go against what I believe technology should be used for.	○	○	○	○	(

AI systems that take part in recruitment decisions are unsettling.

Attention check

Did the AI system you read about explain its recommendation?

- Yes
- No
- Not sure

Demographics

How old are you?

What is your gender

Male Female Prefer not to say

What is your education level?

High School Bachelor Master PhD

What is your experience with AI tools?

None Low Medium High

What is your professional background?

- Human Resources / Recruitment
- Management / Administration
- Other (please specify)
- I have no professional experience yet

What is your country of residence?

Block 12

Do you have any comments or feedback about this study?

Appendix B: *Complete Multiple Regression Results for Fairness*

Variável	*b*	*SE*	*β*	*t*	*p*
AI System Type	0.309	0.103	0.224	2.998	.003
Age	-0.001	0.005	-0.009	-0.113	.910
Male	0.326	0.106	0.235	3.075	.003
Experience with AI Tools	-0.326	0.088	-0.296	-3.707	<.001
Education Level	0.078	0.085	0.076	0.912	.363
Professional Background	-0.133	0.075	-0.139	-1.761	.080
Portugal Resident	-0.123	0.150	-0.061	-0.819	.414

Note: R^2 = .23; adjusted R^2 = .19; $F(7, 140) = 5.95$, $p < .001$; N = 148.
Dependent variable: Fairness.

Appendix C: *PROCESS Model 4 Output for the Parallel Mediation Analysis Including Control Variables*

Model : 4
Y : WOA_w
X : AI_type
M1 : Trust
M2 : Expl

Covariates: Age, Male, Education Level (EduLevel), Experience with AI Tools (ExpAIT), Professional Background (ProfBack), Portugal Resident (PortResi)

N = 138

```
OUTCOME VARIABLE:
 Trust

Model Summary
          R       R-sq        MSE          F        df1        df2          p
      .3710      .1376      .4910     2.9641     7.0000   130.0000      .0065

Model
              coeff         se          t          p       LLCI       ULCI
constant     2.7891      .5112     5.4559      .0000     1.7778     3.8005
AI_type       .3180      .1217     2.6131      .0100      .0773      .5588
Age           .0068      .0053     1.2870      .2004     -.0037      .0173
Male          .2134      .1236     1.7264      .0866     -.0311      .4579
EduLevel      .1477      .1038     1.4227      .1572     -.0577      .3530
ExpAIT        .0805      .1075      .7492      .4551     -.1321      .2932
ProfBack     -.1767      .0894    -1.9778      .0501     -.3535      .0001
PortResi     -.2653      .1792    -1.4802      .1412     -.6199      .0893
```

```
OUTCOME VARIABLE:
 Expl

Model Summary
          R       R-sq        MSE          F        df1        df2          p
      .7114      .5060      .5770    19.0260     7.0000   130.0000      .0000

Model
              coeff         se          t          p       LLCI       ULCI
constant     1.7831      .5542     3.2173      .0016      .6866     2.8795
AI_type      1.4849      .1319    11.2541      .0000     1.2239     1.7459
Age           .0040      .0057      .7030      .4833     -.0073      .0154
Male         -.0674      .1340     -.5028      .6160     -.3324      .1977
EduLevel     -.0425      .1125     -.3781      .7060     -.2652      .1801
ExpAIT        .3755      .1165     3.2228      .0016      .1450      .6061

ProfBack     -.0571      .0969     -.5896      .5565     -.2488      .1345
PortResi      .1918      .1943      .9873      .3254     -.1926      .5763


OUTCOME VARIABLE:
 WOA_w

Model Summary
          R       R-sq        MSE          F        df1        df2          p
      .3625      .1314      .3017     2.1518     9.0000   128.0000      .0296

Model
              coeff         se          t          p       LLCI       ULCI
constant      .5546      .4467     1.2415      .2167     -.3293     1.4384
AI_type       .0281      .1347      .2087      .8350     -.2385      .2947
Trust         .0764      .0738     1.0346      .3028     -.0697      .2225
Expl          .0649      .0681      .9527      .3425     -.0699      .1997
Age          -.0048      .0042    -1.1516      .2517     -.0131      .0034
Male          .1295      .0985     1.3140      .1912     -.0655      .3244
EduLevel      .0845      .0823     1.0274      .3062     -.0783      .2474
ExpAIT       -.1336      .0876    -1.5246      .1298     -.3070      .0398
ProfBack     -.0949      .0711    -1.3351      .1842     -.2356      .0458
PortResi     -.2548      .1431    -1.7800      .0774     -.5380      .0284

************************** TOTAL EFFECT MODEL ****************************
OUTCOME VARIABLE:
 WOA_w

Model Summary
          R       R-sq        MSE          F        df1        df2          p
      .3321      .1103      .3043     2.3025     7.0000   130.0000      .0303

Model
              coeff         se          t          p       LLCI       ULCI
constant      .8834      .4025     2.1949      .0299      .0871     1.6796
AI_type       .1488      .0958     1.5527      .1229     -.0408      .3384
Age          -.0040      .0042     -.9665      .3356     -.0123      .0042
Male          .1414      .0973     1.4533      .1486     -.0511      .3339
EduLevel      .0931      .0817     1.1388      .2569     -.0686      .2547
ExpAIT       -.1031      .0846    -1.2181      .2254     -.2705      .0643
ProfBack     -.1121      .0703    -1.5941      .1134     -.2513      .0270
PortResi     -.2626      .1411    -1.8609      .0650     -.5418      .0166
```

```
************** TOTAL, DIRECT, AND INDIRECT EFFECTS OF X ON Y **************

Total effect of X on Y
     Effect         se          t          p       LLCI       ULCI
      .1488      .0958     1.5527      .1229     -.0408      .3384

Direct effect of X on Y
     Effect         se          t          p       LLCI       ULCI
      .0281      .1347      .2087      .8350     -.2385      .2947

Indirect effect(s) of X on Y:
          Effect     BootSE   BootLLCI   BootULCI
TOTAL      .1207      .0923     -.0585      .3058
Trust      .0243      .0231     -.0213      .0734
Expl       .0964      .0927     -.0827      .2813
```