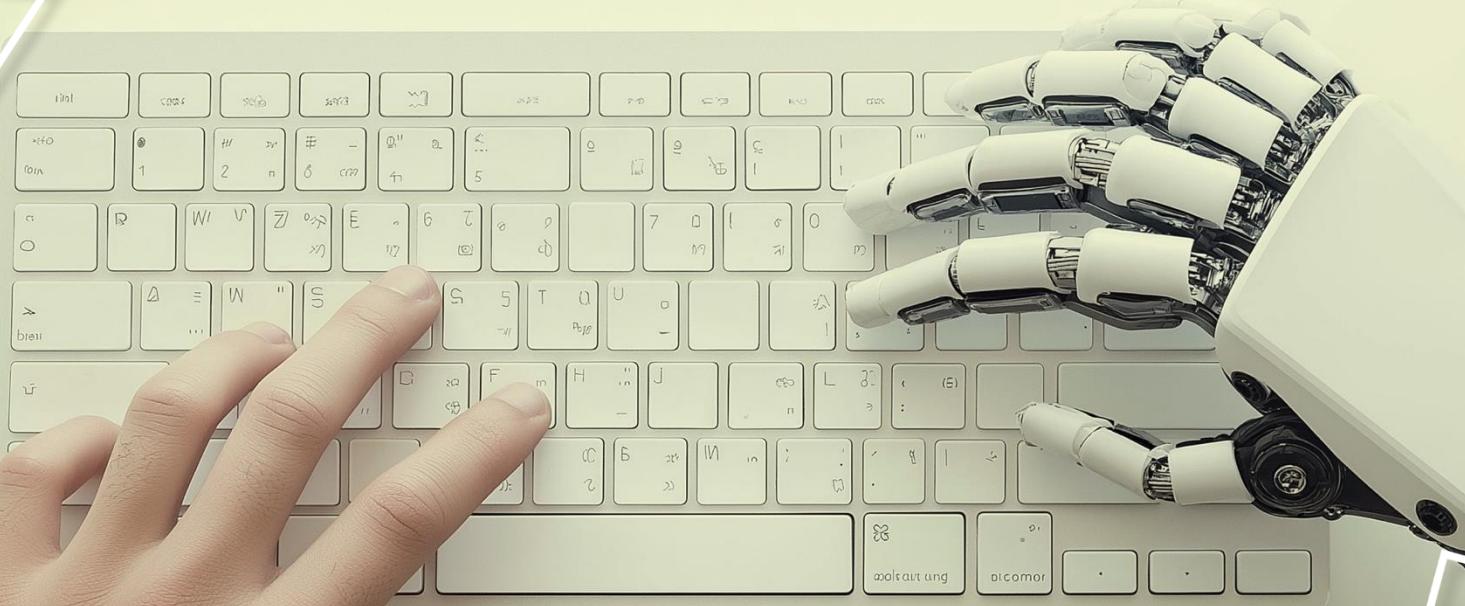




# THE IMPACT OF HUMAN-AI INTERACTION ON DISCRIMINATION

A large case study on human oversight of AI-based decision support systems in lending and hiring scenarios.

EU Policy Lab



2025

RESPONSIBLE AI

HUMAN OVERSIGHT

DISCRIMINATION

Joint  
Research  
Centre

EUR 40136

This document is a publication by the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The contents of this publication do not necessarily reflect the position or opinion of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication. For information on the methodology and quality underlying the data used in this publication for which the source is neither Eurostat nor other Commission services, users should contact the referenced source. The designations employed and the presentation of material on the maps do not imply the expression of any opinion whatsoever on the part of the European Union concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

#### Contact information

EU Policy Lab

Unit S.1 EU Policy Lab: Foresight, Design and Behavioural Insights  
Joint Research Centre, European Commission, Brussels, Belgium  
JRC-FORESIGHT@ec.europa.eu

EU Science Hub

<https://joint-research-centre.ec.europa.eu>

JRC139127

EUR 40136

PDF ISBN 978-92-68-22566-0 ISSN 1831-9424 doi:10.2760/0189570 KJ-01-24-180-EN-N

Luxembourg: Publications Office of the European Union, 2025

© European Union, 2025



The reuse policy of the European Commission documents is implemented by the Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission documents (OJ L 330, 14.12.2011, p. 39). Unless otherwise noted, the reuse of this document is authorised under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence (<https://creativecommons.org/licenses/by/4.0/>). This means that reuse is allowed provided appropriate credit is given and any changes are indicated.

For any use or reproduction of photos or other material that is not owned by the European Union permission must be sought directly from the copyright holders.

Cover page illustration by Curioso.Photography on Adobe Stock

How to cite this report: European Commission: Joint Research Centre, Gaudeul, A., Arrigoni, O., Charisi, V., Escobar Planas, M. and Hupont Torres, I., The Impact of Human-AI Interaction on Discrimination, Publications Office of the European Union, Luxembourg, 2025, <https://data.europa.eu/doi/10.2760/0189570>, JRC139127.

# The Impact of Human-AI Interaction on Discrimination

A large case study on human oversight of AI-based decision support systems in lending and hiring scenarios.

Authors:

Alexia Gaudreul

Ottla Arrigoni

Vasiliki Charisi

Marina Escobar-Planas

Isabelle Hupont

2025

## EU Policy Lab

The EU Policy Lab is a space for cross-disciplinary exploration and innovation in policymaking. We apply collaborative, systemic, and forward-looking approaches to help bringing the scientific knowledge of the Joint Research Centre into EU policy making.

We experiment with the new, the unprecedented and the unknown. We seek to augment our understanding of the present, challenge and reinvent the way we think about the future.

The EU Policy Lab is also a mindset and a way of working together that combines stories and data, anticipation and analysis, imagination, and action. We bring new practical and radical perspectives to tackle complex problems in a collaborative way. Together, we explore, connect, and ideate to create better policies.



[policy-lab.ec.europa.eu](http://policy-lab.ec.europa.eu)

# Contents

Abstract .....	1
Acknowledgements.....	2
Executive summary.....	3
Policy Context .....	5
Key conclusions .....	5
Main findings.....	6
Related and future JRC work .....	7
Quick guide.....	7
1   Introduction.....	8
2   Context.....	10
3   Research questions .....	13
4   Methods.....	16
4.1   A behavioural experiment:.....	16
4.2   Post-experimental qualitative studies .....	25
5   Results of the experiment (Quant).....	33
5.1   Descriptive statistics .....	33
5.2   Analysis of choices.....	39
6   Results of the study's qualitative part (Qual).....	50
6.1   Sample descriptive statistics.....	50
6.2   Analysis of results .....	52
7   Discussion.....	69
7.1   Human and organisation biases (Overseer).....	71
7.2   Oversight of the overriding (Overseer) .....	72
7.3   Mutual checks (Human + AI).....	72
7.4   Outcome feedback and reinforcement learning (Human + AI).....	73
7.5   Outcome monitoring and alignment (Decisions).....	74
8   Conclusion .....	75
References .....	76
List of abbreviations and definitions .....	82
List of figures.....	82
List of tables.....	83
Annexes .....	84
Annex 1. Sample characteristics vs. quotas .....	84
Annex 2. Variables collected .....	87
Annex 3. Preferences of the AI-based DSS.....	99
Annex 4. Preferences and prejudices of the deciders.....	101
Annex 5. Regressions.....	102

## Abstract

This large-scale study assesses the impact of human oversight on countering discrimination in AI-aided decision-making for sensitive tasks. We use a mixed research method approach, in a sequential explanatory design whereby a quantitative experiment with HR and banking professionals in Italy and Germany (N=1411) is followed by qualitative analyses through interviews and workshops with volunteer participants in the experiment, fair AI experts and policymakers. We find that human overseers are equally likely to follow advice from a generic AI that is discriminatory as from an AI that is programmed to be fair. Human oversight does not prevent discrimination when the generic AI is used. Choice when a fair AI is used are less gender biased but are still affected by participants' biases. Interviews with participants show they prioritize their company's interests over fairness and highlights the need for guidance on overriding AI recommendations. Fair AI experts emphasize the need for a comprehensive systemic approach when designing oversight systems.

Keywords: Artificial Intelligence, Human Oversight, Discrimination, Decision Support Systems, Bias, Fairness, Responsible AI.

## Acknowledgements

We thank Marianna Baggio, Anita Braga, and Songül Tolan for their contributions in the early phases of this project. We also thank participants who agreed to be interviewed and take part in group session after the experiment. We finally thank experts who took part in our collaborative speculative workshop: Valeria Adani, Egon L. van den Broek, Raziye Buse Çetin, Filippo Cuttica, Manuel Dietrich, Abdelrahman Hassan, Tim de Jonge, Suhair Khan, Senka Krivic, Christina Melander, and Giada Pistilli.

Marina Escobar-Planas' work was carried out with the support of the Joint Research Centre of the European Commission in the framework of the Collaborative Doctoral Partnership Agreement No.35500.

## Executive summary

This study examines how human oversight can improve fairness in AI-assisted decision-making, particularly in sectors like human resources and credit lending. While AI is increasingly used to support decision-makers, humans still make the final decisions, which helps prevent harm but may also introduce unintended biases. Our findings highlight how human and AI interactions influence decision-making, especially regarding fairness and discrimination.

### Method and Findings:

We conducted experiments with professionals in human resources and credit lending to examine the potential for discrimination in AI-supported decision-making. We used an explanatory sequential design, following the below mixed-methods approach steps:

 We collected data measuring 500 people effort (real-effort task) and trustworthiness (trust game) in the lab and analysed their decisions using machine learning.

 With the data collected, we trained 2 AI-based Decision Support Systems (DSS):

- One optimized for fairness (protecting inputs gender and nationality)
- One optimized for accuracy (a generic AI)

 We recruited 1400 field professionals who chose who to hire and who to lend, and showed them either a model with or without AI recommendation. Professionals made decisions about candidates, rating attributes such as: interview performance, income and education. In the study, we gauged participants' preferences by asking them to choose and to rank the importance of various candidate characteristics.

Quantitative Results: showed that decision-makers were no more likely to follow the fair AI's recommendations than those of the "unfair" generic AI. Gender discrimination against men disappeared when using the fair AI, whereas the generic AI introduced a bias against women. Similarly, the generic AI resulted in discrimination against Italian applicants. The generic AI, which favoured men and Germans, thus influenced choice against women and Italians. The fair AI for its part influences choice to be less discriminatory against men. Fair AI did thus appear to reduce gender discrimination, but decision makers' preferences also played a role. We confirm that the decider's preferences also have an impact, but that individual preferences do not have more of an influence on choice when there is an AI or none. This allays the concern that even fair AI may enable more precise discrimination based on the decider's preferences.

 In the follow-up qualitative study, we conducted semi-structured interviews and small-group workshops with a subset of study participants. The interviews explored participants' real-life experiences with AI, their decision-making processes, perceptions of bias in candidate selection, and their reasoning in the experiments' scenarios.

**Qualitative Results:** Overall, participants held a positive attitude towards AI for professional purposes. They discussed the distinctions between their personal biases, those of their organizations, and broader societal prejudices. They generally prioritized their employer's perspective over challenging organizational norms. Participants believed they were better at assessing situations on a case-by-case basis than AI and expressed hesitance about AI's ability to assess "soft" qualities, such as interview performance. Concerns were raised about the lack of feedback on both their and the AI's performance, which limited their ability to evaluate final decisions. They also emphasized the need for clearer guidance on when to override AI recommendations. While they found the experimental setup relevant to real-life situations, they noted issues with the selected candidate characteristics, their assigned weights, and the format of AI recommendations.

Following the engagement with participants, we organised a participatory design workshop where experts generated ideas on fairness and bias in AI-supported decision-making, addressing six main themes:

- defining algorithmic and human fairness
- translating fairness into practical rules for human-AI collaboration
- regulatory requirements for oversight
- mutual checks between human and AI
- fostering awareness among users and developers
- potential policy directions.

Finally, we invited policymakers to a workshop to reflect on our findings and discuss policy implications. This workshop aimed to examine how our findings could inform practical guidelines for human oversight in AI-supported decision-making. Outcomes included proposals for regulatory guidelines, stakeholder engagement and training initiatives, and methods to monitor and evaluate AI systems.

### Main Conclusions:

Our study illustrates how human and algorithmic biases can intersect rather than cancel each other out. Human oversight, while essential, may not fully correct outcomes from biased AI and may introduce additional biases.

These findings highlight the need for improvements in AI oversight systems, shifting from individual oversight to an integrated system designed to mitigate human bias. Oversight should go beyond individual reviewers or programming fair AI algorithms; it should encompass systemic fairness, involving stakeholders across the AI lifecycle to address both technical and social dimensions. Effective oversight requires guidelines to assist decision-makers in determining when to override AI recommendations, systems that monitor AI-assisted outcomes to identify emerging biases, and mechanisms allowing users to justify overrides. Decision-makers should have access to data on their performance and biases, and AI systems should be regularly evaluated based on user feedback. Experts and policymakers emphasized the need to assess real-world outcomes of AI-human interactions over mere rule compliance. These recommendations aim to enhance the performance, fairness, and acceptability of AI-assisted decision support systems.

## Policy Context

The rapid adoption of AI in decision support systems (DSS) has brought fairness and demographic bias concerns, particularly in high-risk areas such as credit lending and recruitment. These, along with other ethical challenges, have catalysed a regulatory response from EU policymakers, culminating in the adoption of the AI Act, which entered into force in August 2024.

A cornerstone of the AI Act is its requirement to ensure effective human oversight mechanisms for high-risk AI systems—an essential safeguard for promoting the responsible and ethical use of AI through human supervision. The issue of non-discrimination is also deeply embedded in the AI Act, reflecting the EU's commitment to protecting fundamental rights as per Article 21 on non-discrimination in the EU Charter of Fundamental Rights. This is evident in its Whereas 67, which highlights the risks of discrimination arising from AI systems, particularly for vulnerable groups such as ethnic minorities. It emphasizes the critical role of high-quality data governance in preventing discrimination, calling for training, validation and testing datasets that are relevant, representative, and as free of errors and biases as possible. It also warns against feedback loops in AI systems that may amplify existing inequalities. Furthermore, Article 10 obliges providers of high-risk AI systems to implement robust data governance practices, including the examination of datasets for biases and the adoption of measures to detect, prevent and mitigate such biases. These provisions underscore the policy relevance of transparency, human oversight and non-discrimination in AI development.

Building on these foundations, our study explores the concept of human oversight in AI decision-making. We define human oversight broadly, encompassing humans' involvement in AI decision-making, humans' taking decisions using AI suggestions, and human interaction and collaboration with AI. In line with the concept of human oversight in the AI Act, we do not only look at "ex-ante": programming, testing and understanding of AI applications, but especially at "ex-post": monitor, review and influence on AI decisions. We emphasize the importance of continuous monitoring, post-deployment review, and the ability to override AI decisions when necessary. This dual-phase perspective is essential for ensuring oversight mechanisms are both proactive and responsive in addressing real-world challenges.

In the context of the AI Act, our study is particularly relevant for addressing current gaps in implementation, especially concerning bias and discrimination. The provisions on human oversight (Article 14) are foundational but require further operationalization to translate their principles into actionable practices. Our work aims to guide this transition, providing insights that can inform future standards and guidelines for effective human oversight and bias mitigation. Additionally, our findings could play a pivotal role in the development of regulatory sandboxes for AI, as envisaged by the AI Act. Sandboxes offer controlled environments to test and refine AI systems under regulatory supervision, making them ideal platforms to explore the practicalities of human oversight, data governance and anti-discrimination measures. By framing oversight within real-world scenarios and promoting a holistic system-level approach that integrates technical and human considerations, we aim to contribute to the establishment of best practices that are both robust and adaptable. This is essential not only for compliance with the AI Act but also for fostering public trust in AI systems deployed in high-risk domains.

## Key conclusions

The study reveals that AI-supported decision-making systems, when combined with human oversight, can both perpetuate and mitigate biases. Existing policies often assume that human oversight will automatically counteract AI biases. This study overturns that assumption, highlighting that human biases can also influence decision-making processes, especially outcomes, even when AI systems are designed to be fair.

This study thus highlights the need for a multi-faceted approach to AI oversight, integrating technical, organizational, and policy measures. Policymakers must address the dual challenges of AI and human biases by ensuring the monitoring of the outcomes of AI-human collaboration. The existing policy options, when facing implementation, should be reassessed to account for the potential biases introduced by human overseers. This re-assessment should consider the effectiveness of current measures in addressing these biases and explore alternative options for improving oversight.

Potential interventions include intervening with enhanced feedback mechanisms for the combined decision making, improved bias detection tools that are not limited to simply testing the AI for bias, and more effective human-AI collaboration frameworks that allow for complementary human input. While this research improved our understanding some aspects of AI-human interaction, substantial work remains to refine policies and practices for effective oversight of AI systems.

## Main findings

The integration of Artificial Intelligence (AI) in decision-making processes in sectors such as credit lending and recruitment presents both opportunities and significant challenges. This study underscores the complexities and risks associated with AI-supported human decision-making, focusing on biases from both AI and human overseers. The findings provide crucial insights for policymakers aimed at developing robust oversight frameworks to mitigate these risks and ensure fair and non-discriminatory outcomes.

- a. Impact of Organizational Norms on AI Oversight: Overseers often conform to AI biases when these biases align with organizational norms and objectives. This conformity can perpetuate discriminatory practices. Policies could integrate AI oversight with broader anti-discrimination laws and initiatives, ensuring that organizational practices do not inadvertently support biased decision-making. Policymakers would need to create integrated frameworks that address both AI-specific and broader anti-discrimination policies.
- b. Review and Monitoring of Override Decisions: Human overseers sometimes override AI decisions based on their own biases, which can counteract the benefits of fair AI systems. This suggests the need to implement mechanisms to review and monitor override decisions. The outcome of AI-assisted human decisions should be audited to detect and mitigate biases. AI systems should be regularly reviewed on that basis to improve their fairness and reliability.
- c. Critical and Complementary AI-Human Decision-Making: Human overseers value their ability to assess nuanced, context-specific attributes of candidates that AI may not fully capture. This suggests the need to foster complementary AI-human decision-making where AI assists with data processing and humans provide contextual judgment. Transparency in AI systems and explanations for decisions helps in fostering such complementary human input. The role of overseers should not be limited to approving AI decisions or not, but there should be clear guidelines to guide their input so as not to bias outcomes.
- d. Feedback Mechanisms for Continuous Learning: Overseers need feedback to understand whether AI-supported decisions are correct. There is a need for continuous feedback loops between AI systems and human overseers. This would encourage reinforcement learning that combines human and AI feedback to improve decision-making processes.
- e. Outcome Monitoring for Fairness: Human oversight can introduce biases in AI-supported decisions, necessitating robust outcome monitoring. Dynamic and continuous monitoring of AI outcomes would ensure they remain fair over time. Policies that mandate testing AI systems for fairness and reliability should also include testing the outcome of AI-supported decisions. This would ensure that AI systems are fair ex-post, after human intervention, not just ex-ante.

## Related and future JRC work

The EU Policy Lab, as S1, is engaged in the examination of artificial intelligence from a social sciences perspective. It aims to assess and comprehend AI's effects across diverse societal sectors by employing analytical methods that focus on human behaviour, predictive futures, and the interconnected nature of societal systems. Current and future work is focusing on three workstreams:

1. Using AI as a research and communication tool, personalizing and tailoring use cases of AI and experimenting with how it can shape our research practices. For example, we design and establish research practices for Foresight analysis based on AI (Ai4Foresight).
2. Behavioural analysis and experiments to understand the implications of the use of AI on people. Example of specific research questions: What are the potential issues with the use of AI in the short-run (hallucinations, misinformation) and in the long-run (human autonomy)? How do we ensure safe integration of AI into various fields from education to policymaking?
3. Exploring consequences across various areas of the rise in the use of AI, analysing their connections and future implications from a macro systemic point of view. For example: Understanding the contextual factors for further development of AI in research, and innovation; Researching the impact of human oversights on discrimination, linking this with wider considerations about human rights, and understanding the role of institutions and regulations in this respect.

This experiment is part of the workstreams 2 and 3.

## Quick guide

This report offers critical insights for policymakers about the influence of Artificial Intelligence (AI) on human decision-making, particularly within credit lending and recruitment. It aims to provide policymakers with evidence-based recommendations on designing and implementing oversight mechanisms to ensure that AI systems uphold fairness and protect fundamental rights.

Key topics: An analysis of how AI is used to enhance decision-making efficiency and consistency, with potential risks of perpetuating or amplifying human biases.

- AI in Decision Support Systems: Understanding AI's role in improving decision processes and the associated risks.
- Automation Bias: The issue of over-reliance on AI recommendations.
- Algorithm Aversion: Rejecting AI advice due to lack of trust, overconfidence, differences in preferences, and different ways to make and justify decisions.

Research Methodology: Mixed methods research using a sequential explanatory design. In the first phase we collected and analysed quantitative data from an experiment involving a hybrid Human AI decision process. In the second phase we collected and analysed qualitative data from interviews and focus groups with participants in the experiment, to help explain and elaborate on the quantitative results. We led co-design workshops with fair AI experts and policymakers who deal with AI policy at the EC. This allowed us to further our qualitative research and make sense of our results.

Contributions:

Evaluation of how AI impacts decision outcomes and of the effectiveness of human oversight in preventing discrimination, including priorities for considerations and potential interventions.

# 1 Introduction

Artificial Intelligence (AI) is becoming more widely used in all areas of business along with progress in digitalisation (Eurostat 2024). One of its domains of applications is to aid human decision making in high-stakes areas such as credit lending and recruitment. AI is used in Decision Support Systems (DSS) to make faster and more consistent decisions in areas where human decision making is affected by cognitive biases and limitations.

However, the use of AI in DSS raises several issues, one of which is that its decisions may conflict with norms of fairness and may be affected by its own learned machine biases. Those biases often correspond to human ones, which means that AI carries the risk of perpetuating or even amplifying discrimination against groups of people, based for example on nationality, gender or race (Kordzadeh and Ghasemaghaei 2022; Mehrabi et al. 2021; Mitchell et al. 2021; Tolan 2019; Vlasceanu and Amodio 2022). Such discrimination is illegal under Title 3 of the Charter of Fundamental Rights of the European Union (2012).

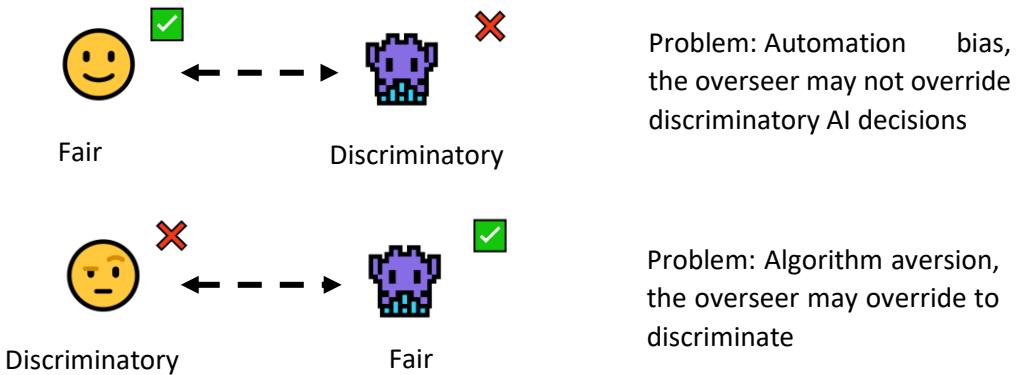
The European Union has therefore dealt with this issue, and others related to digitalisation, with a succession of legislative measures. The General Data Protection Regulation gives the right not to be subject to a decision based solely on automated processing (Article 22). The AI Act requires human oversight of AI to prevent or minimize risks to fundamental rights. Article 14 requires that AI systems are designed and developed such that they can be effectively overseen by natural persons. Among other provisions, the human overseer must be able to fully understand and interpret the AI system's output<sup>1</sup> and must have the option not to use it. The Directive on Platform Work also addresses risks of discrimination by providing for the right to get algorithmic decisions reviewed by a human (Article 10). In European, but also in international law, human oversight is thus advocated as a solution against the risks of increasing reliance on algorithmic tools (Koulu 2020). Human overseers are supposed to increase the accuracy and safety of AI systems, uphold human values, and build trust in the technology (Laux 2023). Human oversight is one of a range of different measures that are supposed to ensure that human values are reflected in decision-making and consequently human fundamental rights, such as agency, are protected.

Our goal in this work is to investigate the role of humans in oversight systems, and how those systems must be designed to prevent discriminatory outcomes from the use of AI. We investigate two behavioural biases that can make human oversight ineffective or even counterproductive. The first is an automation bias, which is specifically mentioned in the AI Act, whereby people automatically rely on the AI system and do not challenge it. This is a problem if the AI is discriminatory (Figure 1). The other bias is known as algorithm aversion (Mahmud et al. 2022), whereby people reject algorithmic decisions in favour of what they think is best. This is a problem if the decision-maker is discriminatory.

---

<sup>1</sup> See (Panigutti et al. 2023) for more details.

Figure 1 Exploring the combination of human and AI biases.



Source: Own elaboration

We explore the effect of those two contrasting biases by eliciting preferences of decision-makers and giving them advice from either unbiased or biased AIs. We measure the rate at which they follow AI recommendations and relate this with their own preferences. This allows us to determine whether giving unbiased AI advice makes human decisions less discriminatory than unaided human decisions, and conversely whether biased AI advice makes human decisions more discriminatory. Our study is thus situated in the general field of investigation of human-AI complementarity, but innovates compared to the usual study that focuses on comparing the *performances* of AI on its own, humans on their own, and AI-humans teams (Patrick Hemmer et al. 2021). We focus instead on a comparison of the *levels of discrimination* that result when AI is left to make decisions on its own, when humans are on their own, and when humans are provided with support from AI to make decisions.

We show in this study that users are influenced in the direction of discrimination suggested by a discriminatory AI, and that users override suggestions made by a fair AI to fit their own discriminatory preferences. Neither do fair users prevent discriminatory outcomes of unfair AI, nor does fair AI prevent users from making discriminatory decisions.

We follow on this study with interviews and workshops with participants in our experiment to investigate further their reaction to AI support and contextualise their decisions. With fair AI experts, we then analyse the context of the study and the ethical considerations at play in decisions and draw lessons from this large-scale study about the proper way to implement human oversight systems to avoid discriminatory outcomes.

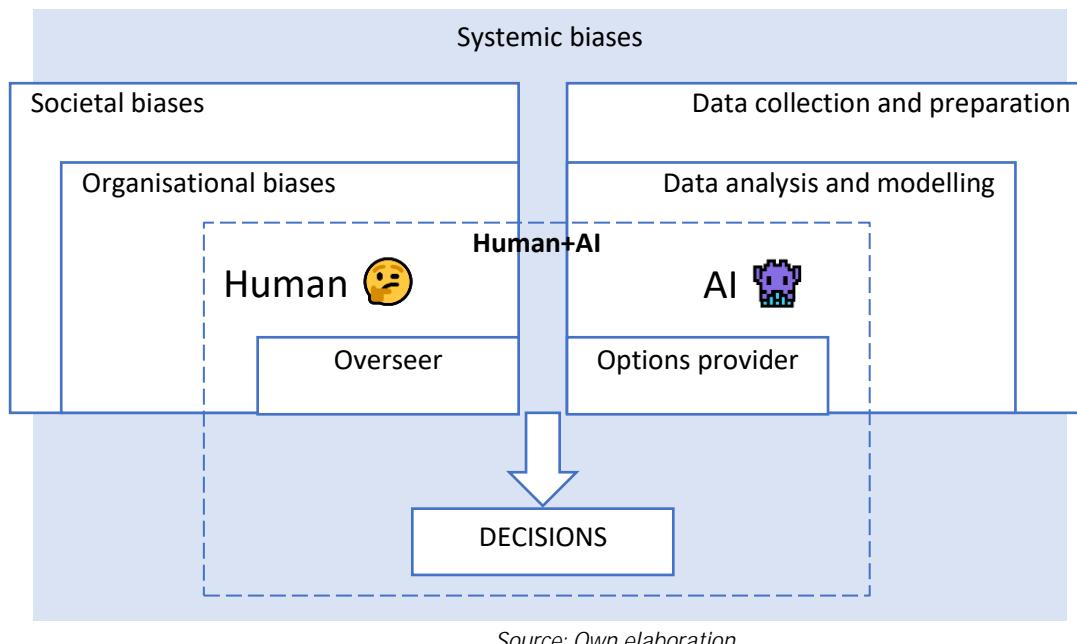
## 2 Context

We introduce in this part some background to our study: how AI is being used to assist human decisions, its effects in terms of discrimination, and the measures that have been taken to prevent such discrimination, including the requirement to maintain human oversight of AI systems.

Domains of application: AI is being used to assist human decision-making in sensitive domains such as hiring, lending, deciding on medical treatments, and sentencing. We consider in this study the use of AI to select applicants in the domain of human resources and banking. In human resources, AI helps in identifying prospective recruits, checking their references, gathering information about them from different sources, predicting job performance, and streamline the entire recruitment lifecycle, from job posting to candidate selection, including interviews which can be led by AI powered chatbots (IBM Consulting 2023). In banking, AI can for example help determine an applicant's creditworthiness, automate loan approval, run risk assessment and credit scoring, detect fraud, make personalized loan offers. Chatbots can help in processing loan applications and interviewing applicants (Lee 2023).

AI and Human decision making: In this work, we concentrate on the specific issue of how humans and AI interact in making decisions that impact other humans, and we focus even more precisely about biases in those decisions and where they come from (Figure 2).

Figure 2 Human and AI decisional background and their interplay.



The background and sources of biases in AI and human decision making in such a setting are quite different: Biases that occur in human decision making have systemic, societal, organisation or individual grounds. They manifest in various forms, such as racial profiling, unequal pay, and limited opportunities for advancement. Systemic biases are those that are inherent in the structures and processes of institutions, and can manifest in various forms, such as policies, practices and attitudes that produce chronic adverse outcomes for specific populations. Societal biases are those

that are ingrained in the culture and norms of a society, such as gender roles and racial stereotypes. Organizational biases are those that are inherent in the policies, practices, and culture of an organization, including decision-making processes that are biased towards certain groups, or organizational cultures of exclusion or marginalization. Finally, individual biases are personal beliefs and attitudes based on personal experience that can influence behaviour and decision-making. They can be explicit or implicit, meaning they may not even be conscious and accessible to reason.

The biases that are embedded in AI cannot be fully separated from human biases, as AI systems are established by humans and will thus reflect their biases. However, the systems that support AI applications are quite distinct and different from the ones that support and guide human decisions. AI is based on big data and machine learning, which it uses to inform decisions. When having to make decisions about a human, the AI system will consider a range of individual characteristics such as income, wealth, age, gender, nationality, education, occupation, residence, and so on. This can potentially lead to discrimination based on protected characteristics. This can arise if the data used to train the AI is affected by pre-existing biases or is not representative of the population. It happens also if the design of the algorithm is itself biased by preconceptions in terms of the choice of features to consider or the formulation of the problem to solve. Correlation between characteristics, such as race and residence, can also give rise to proxy discrimination (Celi et al. 2022; Ferrara 2024; Schwartz et al. 2022).

**Discriminatory outcomes from the use of AI:** There is a large set of evidence of discrimination against groups of people based on their gender, nationality, ethnicity, and other factors. (Zick, Küpper, and Hövermann 2011) is just one report focusing on EU countries. This discrimination affects many domains such as access to health services, employment, or education. Most of this discrimination is due to human biases and prejudices, and the systems that support them. The role of AI in sustaining such discrimination has only emerged more recently in several scandals that outlined the discriminatory outcomes resulting from the use of automated decision systems, and their negative impact on individuals and groups in society. This was the case for example when rating a defendant's risk of future crime (Mattu 2016), when deciding who is a high-risk patient needing extra health care (Ledford 2019), who gets targeted for suspension and investigation of childcare benefits (Amnesty International 2021), or for investigation of social security fraud (France Info 2022).

**Fighting discrimination:** A first and most obvious way to prevent and fight AI discrimination is to develop fairer AI based decision-making processes. Much effort has thus been devoted to documenting and correcting bias in AI output (Barocas, Hardt, and Narayanan 2023; Mehrabi et al. 2021; Zliobaite 2017). However, technology cannot solve such a complex social problems as discrimination because complex issues cannot be reduced to simple engineering problems (Morozov 2013). This is why a community of researchers from social and behavioural sciences has emerged to work towards understanding the impact that algorithmic DSS have on human decision making and consequently on human society (Gerdon et al. 2022). The need for AI oversight systems has emerged along with the need to improve the quality of AI. Those oversight systems are meant to uphold values and principles of ethical AI (Reinecke et al. 2023; Slavkovik 2023; Tsamados et al. 2021). Those principles include transparency, privacy, accountability, fairness, and contestability whenever humans are subjected to AI decisions (Amnesty International and Access Now 2018; A. A. Khan et al. 2022; UNESCO 2021). One of their aims, beyond maintaining accuracy and safety of AI systems, is to ensure that AI systems do not infringe on fundamental rights, including the right not to be discriminated based on protected characteristics.

Human oversight: Maintaining human oversight over AI has been proposed as a way to deal with many of the ethical issues raised by the use of AI in a wide range of applications. Human oversight relies in large part on human intervention at various stages when setting up, using, and maintaining AI systems. Human oversight is dependent on a system of technical and procedural safeguards to be effective. Every step at which human oversight occurs must be supported by institutional and AI design decisions.

The different steps of human oversight can be classified as either *ex-ante*, by reviewing and evaluating AI systems before they are deployed or used, assessing their potential risks and impacts and ensuring that they align with human values and ethical standards, or *ex-post*, meaning either reviewing AI suggestions before implementing them, or reviewing AI decisions if they are appealed or lead to issues (Maxwell 2023).

Ex-post human oversight itself can be classified depending on whether a human is in the loop, with a human actively involved in the decision-making process, often intervening to correct or modify the output of the AI system, on the loop, with a human monitoring the AI system's performance and stopping it if necessary, or even out of the loop, with minimal human intervention beyond deciding to initiate the use of the AI system. In all those cases, a human has the ultimate authority and responsibility over the AI system.

### 3 Research questions

In this work, we consider the effectiveness of the combination of ex-ante “system” oversight, i.e. making sure the AI is fair, and ex-post “individual” oversight, i.e. allowing decision makers to override AI decisions. We consider whether biased overriding during ex-post oversight may not negate the benefits of ex-ante oversight, and whether ex-post oversight can reduce the impact of a failure to perform proper ex-ante oversight.

As mentioned in the introduction, the main issue with human oversight is to balance trust and control. Providing a theoretically unbiased AI is only going to translate in less biased decisions if users trust it to make decisions on their behalf. Conversely, users can prevent biased AI decisions only if they maintain their ability and willingness to understand and question AI.<sup>2</sup>

This part therefore discusses the issue of determining appropriate reliance on AI systems (Schemmer et al. 2022). There can be either over or under reliance on AI systems, meaning AI systems may be relied on when they should not be, or not relied on when they should be. For example, an individual may choose to rely on a system that performs less well than they would on their own or choose not to rely on a system when that system would actually do better than they would. There can be both over and under reliance for the same system, whereby for example the AI is good at a subset of tasks that humans are bad at, but the human intervenes at cross-purpose, thus resulting in the worst of both worlds. Reliance behaviour thus affects accuracy in AI-assisted decision-making. Humans and AI systems must work together to make decisions that leverage their respective strengths and weaknesses effectively (Schoeffer et al. 2023).

A whole strand of research has therefore tried to understand why and when humans make mistakes in their reliance on AI systems, and how to remedy this. Algorithm aversion is the term used when people ignore advice from AI even when following that advice would lead to “better” decisions than what they decide on their own (Dietvorst, Simmons, and Massey 2015).<sup>3</sup> This can be due to the wish to maintain a sense of agency, or to not understanding the logic behind AI advice. Automation bias is the term used when humans follow AI advice even when it is inadequate (Parasuraman and Manzey 2010). This can be due to the perception that automated systems are more reliable or accurate than human judgment.<sup>4</sup>

Some of the research on AI reliance has already focused on how human intervention interacts with AI outcomes with regards to discrimination. (Ghasemaghaei and Kordzadeh 2024) find that the reason why human oversight may not reduce bias is that decision-makers do not necessarily experience guilt when adhering to biased algorithms. This phenomenon is interpreted through the lens of the obedience to authority theory, which suggests that the responsibility for unethical actions shifts from the individual executing the behaviour to the authority figure directing it. On the other hand, being exposed to a fair DSS may reduce discriminatory beliefs of decision-makers. (Avery, Leibbrandt, and Vecchi 2023) show how using AI can help close the gender bias in hiring both in terms of supply and demand, by correcting beliefs of evaluators and because AI is perceived as fairer, thus increasing the motivation of females to apply. (Jussupow et al. 2021) gives a less rosy

---

<sup>2</sup> This is why the AI act requires that humans “be able to interpret the high-risk AI system's output”. Explainable AI is not the only mean for this, and the AI Act does not mandate explainable AI techniques. What it mandates is that the overseer can understand (but not necessarily interpret, as most AI models are opaque) an AI system's inputs and outputs (Panigutti et al. 2023).

<sup>3</sup> Another definition of algorithm aversion is when users are more likely to follow advice coming from a human than from an AI (Morewedge 2022).

<sup>4</sup> A popular illustration of this phenomenon is blind trust in GPS navigation systems, which has led some drivers into dangerous and/or comical situations (Hansen 2013).

assessment, whereby only evaluators who are aware of gender stereotypes in society choose to rely less on a gender-biased AI.

Beyond willingness to follow fair or biased algorithms, another strand of literature underlines the impact of selective reliance on algorithms depending on whether it support pre-existing beliefs. (Selten, Robeert, and Grimmelikhuijsen 2023) show how police officers selectively follow AI advice depending on whether it matches their intuitive professional judgment. Humans may thus choose to follow an algorithm only when it aligns with their personal ideals or beliefs (Chugunova and Luhani 2022; Wang et al. 2023). This “motivated reasoning” (Kunda 1990) leads decision-makers to assign greater weight to AI outputs that confirm their biases, while contesting or discounting AI advice that contradicts their preconceptions (Alon-Barkat and Busuioc 2023). Another reason for not adhering to a “fair” AI recommendation is that the outcomes optimized for by the AI system may not align with the goals and preferences of the human decision-makers using the system. When there is a mismatch, the human may be more inclined to selectively adhere to the AI advice that matches their own desired outcomes, even if it is less accurate or fair (Guerdan et al. 2022).<sup>5</sup>

In this paper, we further illustrate the issue of selective reliance, that is, how the preferences of users in terms of outcomes impact their decision whether to follow or reject AI advice. We also advance research on this topic by considering whether AI recommendations can amplify the impact of discriminatory tendencies as outlined in (Khan, 2023). Let's say there's an AI designed to be fair when deciding who gets a loan. It doesn't consider a person's gender at all, just other important information. Now, imagine there's someone using this AI who thinks the same way about those other factors but also believes that one gender should be favoured over the other. This person might use the AI's advice but then change the final decision in some cases, when the result is close, based on gender. This could actually lead to more targeted discrimination. Indeed, if the person wasn't using the AI, they might not be as good at weighing all the other information, and their bias towards a certain gender wouldn't be as precisely applied.

To illustrate the argument further, consider the case of two applicants, one male and one female, for one job requiring them to produce one good of quality  $q$  (Figure 3). Suppose the female applicant, if hired, will turn out to produce quality  $q_F = 7$  and the male applicant will turn out to produce quality  $q_M = 6$ . If this was known and the hiring process was fair (meritocratic), then the female applicant ought to be hired. Let us now consider what happens with and without AI support when the decider is biased against female applicants.

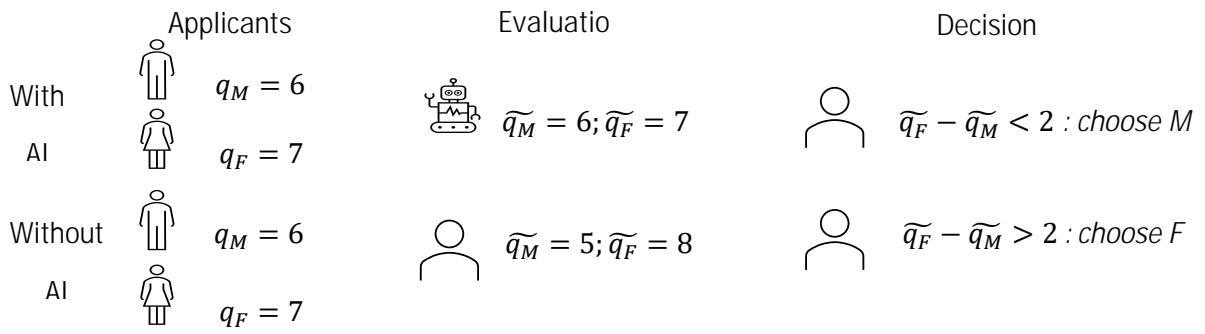
**Decision with AI support:** Suppose an AI can perfectly determine the quality that will be produced by an applicant. The AI tells the decider that  $\widetilde{q}_M = 6$  for the male applicant and  $\widetilde{q}_F = 7$  for the female applicant. Suppose however that the decider dislikes female applicants and is ready to trade quality to obtain a worker of their preferred gender. For example, the decider chooses male applicants as long as  $\widetilde{q}_F - \widetilde{q}_M < 2$ . Then, in our case, the male applicant gets the job.

**Decision without AI support:** Suppose now there is no AI, and the decider can only make noisy evaluations of the “true” quality. It may then happen that the decider overestimates the difference in quality between the two applicants, such that for example the decider estimates  $\widetilde{q}_M = 5$  and  $\widetilde{q}_F = 8$ . In this case, randomness has given a chance to the female applicant.

---

<sup>5</sup> A distinct line of research looks into whether AI can change user preferences, so that being exposed to and getting advice from AI systems may corrupt human morals, for example because people would accept and adopt amoral AI reasoning (Köbis, Bonnefon, and Rahwan 2021; Krügel, Ostermaier, and Uhl 2023; Leib et al. 2021). Those issues and debates have led to calls for integrating human moral principles into AI and for monitoring AI for unmoral behaviour.

Figure 3 Schematic process of evaluation and decision, with and without AI



Source: Own elaboration

This type of argument can be generalized as outlined in (Khan, 2023) to show that under some conditions, AI support can indeed result in giving less chances to a discriminated group even when the AI was programmed not to be biased. We will test for evidence of this possible amplifying effect by testing not only whether discriminatory preferences of the deciders impact affect decisions, but also whether this impact is more pronounced when there is AI support vs. when there is none.

## 4 Methods

This research employs a sequential explanatory design, which is a type of mixed methods research that involves collecting and analysing quantitative data first, followed by collecting and analysing qualitative data to help explain or elaborate on the quantitative results (Fetters, Curry, and Creswell 2013; Creswell 2021). In a first phase we collected and analysed quantitative data from an experiment involving a hybrid Human AI decision process. In a second phase we collected and analysed qualitative data from interviews and focus groups with participants in the experiment, to help explain and elaborate on the quantitative results.

This sequential explanatory design is particularly useful to explore complex situations where quantitative data alone may not provide sufficient context or understanding. In our context, the quantitative data is derived from actual behaviour by people who are called on to make decisions that may be discriminatory. Discrimination is a subject that is typically difficult to explore based on people's self-reports, as most people are not aware or do not want to admit they are discriminating. This is why it was useful in a first step to examine actual levels of discrimination, with or without AI support. On the other hand, the qualitative phase, which is derived from interviews with participants in the experiment, reveal nuances and factors that quantitative data may have overlooked. We cannot expect participants to necessarily be aware of how the AI influenced them or of how far their behaviour was discriminatory. On the other hand, we can ask them about their subjective experience of the experiment, how it related to their own work context, and how they think about the types of issues that are our subject matter. We are thus able to integrate participants perspectives and experiences to better understand the overall topic.

We furthered the qualitative research and sense-making with a co-design workshops with experts and one with policymakers who deal with AI policy at the EC. This allows us to better situate our research and findings in the general field of fair AI, and to generate and express practical insights and research priorities for the future.

### 4.1 A behavioural experiment:

We ran an online performance-based incentivized experiment that mimicked the employer-employee and the lender-borrower relationship. Our experimental study, while still happening in the rather artificial environment of an “online laboratory”, did confront real professionals with decision scenarios that are relatively similar to their real-world professional duties. Those decisions furthermore had real monetary consequences, as applicants receive money only if selected, and deciders make more money when selecting the “best” applicants (in our case, those who repay more of their loan or perform better in an intellectual task).

The experiment was programmed and administered by Ipsos European Public Affairs based on a design provided by the authors.

The full script of the experiment is available as material on the project’s OSF website at <https://osf.io/mhd7r/>, along with a list of the variables collected, data and code for the analysis (in the R statistical language). The experiment, hypotheses and plan of analysis were pre-registered on the OSF registries website at <https://osf.io/5mz3s/>.

Respondents were informed that their responses would remain anonymous and their data private and only used for the purpose of the research. The introduction to the survey also included a link to the Privacy Policy (containing information on data processing, e.g. retention period, how personal

data will be used/processed, whether and how personal data will be shared with third parties, etc.), as well as a question asking participants to give consent for survey participation.

The running of the experiment required us to go through three steps 1) collecting data on the behaviour of participants in the role of applicants for loans and jobs 2) predicting the performance of applicants with a machine learning model<sup>6</sup> 3) asking participants in the role of deciders to choose among applicants.

Those three steps are necessary to ensure that deciders make decisions for real applicants whose data was really analysed using machine learning. In doing this, we differ from other experiments that present hypothetical scenarios to participants. The issue with hypothetical scenarios is that if they are presented as hypothetical, then deciders may not care what decision they make, and if they are presented as real, then the experimenter lies to the participant. This goes against a foundational ethical norm in economic experiments, the principle of no deception, that emphasizes the importance of honesty and transparency in the experimental process. This principle is crucial for maintaining the integrity of experimental economics and ensuring that the data collected is reliable and valid (Charness, Samek, and van de Ven 2022; Krawczyk 2019; Ortmann and Hertwig 2002).

We outline the three steps more in detail in the following.

#### 4.1.1 Applicants:

528 participants in the role of applicants were recruited among the general population between 18 and 65 years old in both Italy (N=274) and Germany (N=254) on the 16<sup>th</sup> and 17<sup>th</sup> of February 2023.<sup>6</sup> Respondents were drawn from Ipsos' Online Access Panels by means of stratified random probability sampling, based on the available profile data – age, gender, and geographic region (NUTS 1) – and pre-defined sub-sample sizes (i.e. quota) based on official population statistics provided by Eurostat. The final sample corresponds closely to those quotas in terms of age category, gender, and regions in each country (Annex 1).

Participants then performed a real-effort task (Charness, Gneezy, and Henderson 2018) and made a decision in a trust game (Berg, Dickhaut, and McCabe 1995). They also responded to a questionnaire.

The real-effort task consisted in computing a series of sums of 4 entire numbers between 1 and 9, such as for example 2+4+3+7, over 5 minutes. They were told they would get a score equal to the number of sums they added up correctly during those 5 minutes. They were also told that, if hired by an employer (a participant in the subsequent HR decider experiment), they would earn a salary of 100 points (corresponding to 4.3€) and their employer would earn points depending on their score, so the higher their score, the more points their employer would earn. Participants were asked to correctly answer two understanding questions and could do a practice round. On average, participants managed to do 61 sums correctly in 5 minutes, with a range from 18 to 120 (the maximum possible), and a median of 58.

The trust game consisted in telling participants that if given a loan from a banker (a participant in the subsequent lending experiment), then they would receive 100 points which they would then invest in a project that earned them 300 points. They were then asked to decide how many points to give back to the banker and told they could repay anything between 0 points and 300 points, as they wished. Participants were asked to correctly answer two understanding questions and could

---

<sup>6</sup> 967 panelists entered the survey of whom 528 completed it, 315 respondents were screened out based on questions checking their understanding of the experiment, and 124 respondents entered the experiment but did not complete it (quit). 3 more participants were excluded based on systematic "Don't know/Prefer not to answer" responses to survey questions with that option.

experiment how different decisions translated into payoffs for them and the banker. On average, participants paid back 108 points, with a range from 0 to 300 and a median of 100.

The questionnaire collected the age, gender, country, region, level of education, nationality, occupation, sector of employment, monthly income and social class (Annex 2). In addition to this, we asked participants questions on their perception of the effort task (Q2\_1 to Q2\_5), which we summarized into an interview score,<sup>7</sup> one question on what they did in the trust game (Q3), questions to evaluate their social attitudes, inspired by the SOEP (risk, trust, competition, fairness, locus of control) (D11\_1 to D11\_5), and a short form of the Big 5 personality questionnaire (D12\_1 to D12\_15) (Lang et al. 2011).

#### 4.1.2 AI-based DSS:

We developed an AI-based Decision Support System (DSS) utilizing the 528 participants' data derived from the previous real-effort task and trust game. The real-effort task and the trust game were structured around two scenarios: a "hiring" scenario and a "banking" scenario, respectively, where participants engaged in roles either as job or loan applicants. The primary objective was to construct a set of AI-based prediction models capable of assisting human deciders in the subsequent decision-making experiments (c.f. section 4.1.3). To this end, we trained four Random Forest models (Parmar, Katariya, and Patel 2019): two aimed at providing support for the "hiring" scenario (one "fair" and one "generic") and two for the banking scenario (one "fair" and one "generic"), each generating a binary output (yes/no decision) based on a set of input variables.

The "fair" models were implemented to make predictions based on non-sensitive personal data, namely: age, level of education, monthly income, and interview score. Conversely, the "generic" models included two additional sensitive or "protected" inputs: gender and nationality. This generic model exhibited discrimination based on those characteristics, meaning they entered into play in the rating of applicants. These different model designs allowed to explore in subsequent experiments the automation bias effect on human deciders when sensitive attributes have an impact on the DSS vs when they do not.

Note that despite having collected a broader range of data from participants (e.g. personality traits, social traits of applicants), we selected only a subset of those variables to align the AI-based decision-making process with that of human deciders. That is, the models' inputs were limited to only those attributes that were directly provided to deciders in the subsequent experiments. Human evaluators thus had access to all information available to the DSS.

The models were trained with discretised input and output variables. The original output of the real-effort task and the trust game were the points earned by participants (discrete values), which we transformed into a binary outcome, which was necessary to apply the fair measures needed for the development of the fair model. This process involved setting a threshold leaving the top 30% of scores as "yes" and the remaining 70% as "no". From the 528 samples collected, this rule resulted in specific thresholds of 69 points for the "hiring" scenario and 125 points for the "banking" scenario. Input variables of continuous nature were similarly discretised into a set of bins, namely: age (5 bins), monthly income (5 bins) and interview score (4 bins).

The Random Forest models were trained using the Scikit-Learn Python library (Scikit-Learn, 2024), employing a 5-fold cross-validation strategy over the 528 samples collected from the real-effort task and the trust game. The accuracy metrics obtained for each of the four models is shown in Table 1.

---

<sup>7</sup> The interview score is  $(Q2\_1+Q2\_2+(5-Q2\_3)+(5-Q2\_4)+Q2\_5)/5$ .

Table 1 Accuracy metrics for each DSS scenario

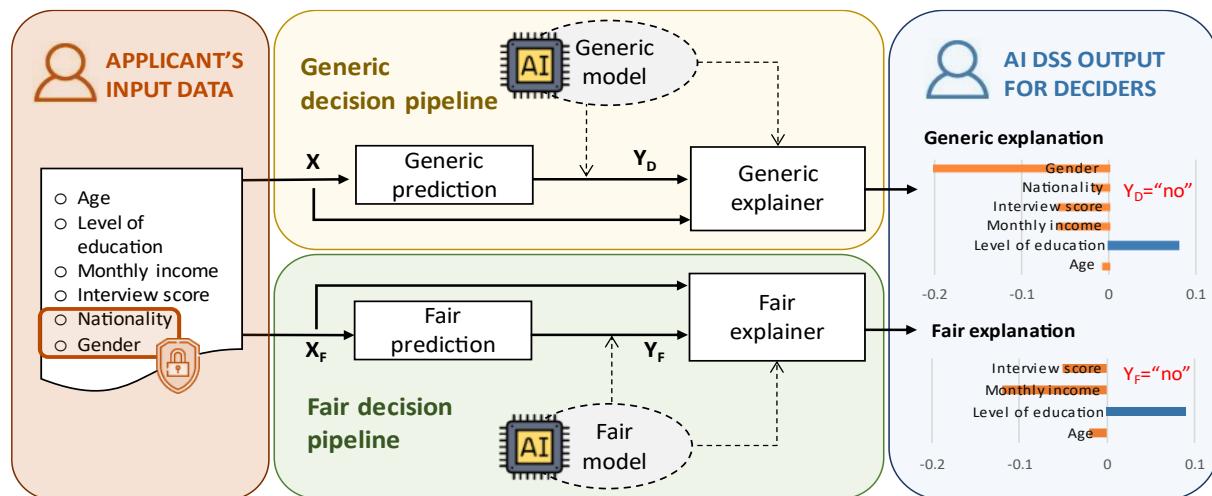
Scenario	Accuracy	Precision	Recall	F1 score
Banking discriminatory	0.917	0.973	0.915	0.943
Banking fair	0.845	0.933	0.860	0.895
Hiring discriminatory	0.913	0.963	0.920	0.941
Hiring fair	0.845	0.947	0.852	0.900

Source: Own analysis

In addition to the yes/no output, our methodology leveraged eXplainable AI (XAI) techniques to provide insights –in the form of positive and negative numerical weights– into how each input variable influenced the model's decision. We used the popular Local Interpretable Model-agnostic Explanations (LIME) XAI technique (M. T. Ribeiro, Singh, and Guestrin 2016), by integrating the LIME Python library into our code (M. T. C. Ribeiro [2016] 2024). This approach aims to foster transparency and an understanding of the DSS by the deciders, who were tasked with making the final hiring or loan granting decisions based on the model's predictions and explanations.

Figure 4 summarizes the computational pipelines followed by the DSS to make “generic” and “fair” predictions. The top pipeline depicts the “generic” one, where inputs to the DSS are  $X=\{\text{age, level of education, monthly income, interview score, nationality, gender}\}$  in their discretised form. Then, the pre-trained “generic” model is used to obtain the “generic” prediction  $Y_D$ , that can be either  $Y_D=\{\text{yes}\}$  (i.e., grant the loan or hire the applicant) or  $Y_D=\{\text{no}\}$  (i.e., deny the loan or not to hire the applicant). Finally, the explainer, based on the actual inputs  $X$ , the pre-trained “generic” model and decision  $Y_D$ , computes explanations providing a positive/negative weight to each input variable according to its influence on the final decision. For instance, it can be seen from exemplar “generic” explanations in Figure 4 that the gender of the applicant has had the highest negative influence on the final decision. This proving that the generic model is indeed in fact discriminatory. The bottom pipeline represents the “fair” decision process. The difference is that inputs do not contain “protected” attributes gender and nationality, i.e.  $X_F=\{\text{age, level of education, monthly income, interview score}\}$ . The pre-trained model used in this case to make the prediction  $Y_F=\{\text{yes}\}$  or  $\{\text{no}\}$  is the “fair” one, which is used together with  $X_F$  inputs by the explainer to obtain explanations. As a result, nationality and gender variables do not influence at all (weight zero) the final “fair” explanations.

Figure 4 Pipelines followed to obtain AI-based Decision Support System's outputs plus explanations for deciders. Top pipeline corresponds to the generation of “generic” outputs and explanations; Bottom pipeline represents the process to obtain “fair” outputs



Source: Own elaboration

When developing our DSS, we faced several challenges that led to a series of implementation decisions we would like to discuss in order to be fully transparent with the algorithmic limitations of our AI-based system. Given the relatively small size of our dataset (528 samples), we opted for a classic Random Forest classifier over a more complex deep learning model. The reason is that deep learning models require larger datasets to effectively learn the underlying patterns without significant overfitting, even when techniques like data augmentation are used. Models like Random Forest have shown to be well-suited to achieve satisfactory performance in smaller datasets, being less prone to overfitting (Grinsztajn, Oyallon, and Varoquaux 2022). Moreover, classic machine learning models rely on hand-crafted features and, as such, Random Forest was able to perform very well on our tabular data. On the other hand, we used a 5-fold cross-validation strategy but did not perform evaluation on a distinct test set. The reason was that we wanted to maximize the use of our data for both training and validation, providing a more robust estimate of model performance than a single train-test split. We believe this approach is reasonable given the dataset size, but comes with some caveats including potential overestimation of the model stability and generalization capabilities with respect to unseen data.

Regarding the implementation of the “fair” decision pipeline, we initially aimed to use the AI Fairness 360 toolkit (Bellamy et al. 2018), providing bias mitigation algorithms for datasets and models. Our focus was on safeguarding against biases related to both gender and nationality, for which we employed the Reweighting pre-processing algorithm (Kamiran and Calders 2012). This algorithm adjusts the weight assigned to each training sample as a penalisation term to counteract biases associated with “protected” attributes. However, our trials revealed challenges in effectively mitigating biases for both “protected” variables simultaneously. Analysis indicated that the “banking” scenario exhibited significant gender bias, whereas the “hiring” scenario was more affected by nationality bias. This discrepancy led to scenarios where one of these variables unduly influenced the model’s decisions, contrary to our objectives of fairness. Consequently, we opted for a more straightforward approach to ensure fairness: excluding gender and nationality from the model’s inputs. This decision, as documented in Table 1, resulted in a performance reduction for the “fair” models (e.g. decrease of accuracy from 0.917 to 0.845 in the “banking” scenario). Nonetheless, this outcome aligns with findings from other state-of-the-art research, underscoring the complex trade-off between ethical considerations and model performance in AI development (Li, Wu, and Su 2023).

In any case, while acknowledging the aforementioned limitations, it is important to note that our primary focus was not on maximising the accuracy of the models but rather on understanding how deciders interact with and are influenced by the outputs of the DSS. Our commitment to utilizing real data and genuine machine learning models aimed to preserve the integrity of the decision-making process, avoiding the potential pitfalls of relying on synthetic data or fabricated predictions.

#### 4.1.3 Deciders:

We recruited 1411 Human Resources and Banking professionals in Italy and Germany between the 24<sup>th</sup> of June and the 30<sup>th</sup> of August 2023.<sup>8</sup> Participants were randomly drawn from B2B panels of HR and finance professionals based on the available profile data (occupation, age, gender, and region). For this sample, however, no official statistics for age, gender, and region for the respective target populations of active professionals (i.e. human resource management and retail banking) are available to set quota; thus no hard quota were applied. The samples were monitored to approach a

---

<sup>8</sup> In HR, 1,340 panellist entered the survey out of which 427 were screened out with questions checking their understanding of the experiment, 174 abandoned the survey before completing it (quit), and a total of 754 completed the survey. In banking, 1,512 panellists entered the survey, out of which 558 respondents were screened out with questions checking their understanding of the experiment, 190 quit the survey, and 764 completed it. The completed surveys were subject to a further quality check by removing participants who spent less than 5mn completing the experiment, and those who spent more than 30mn (hiring experiment) or 44mn (lending experiment). Those limits were based on times spent by the fastest and slowest 5% of participants in a pilot with 41 participants in HR and banking in Germany.

balanced spread for age, gender, and geographic region. The final sample in terms of age category, gender and regions in each country is described in Annex 1.

We had three treatments, varying whether deciders decided on their own, with recommendations of a fair AI, or with recommendations of a discriminatory AI. Table 2 shows the distribution of deciders by country, background, and treatments.

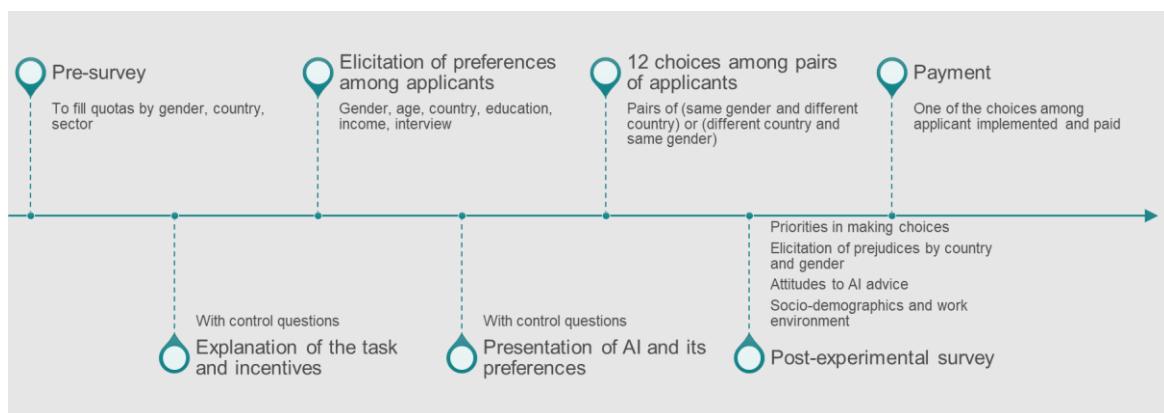
Table 2 Deciders' sample distribution by sector, country, and treatment

AI	Banking		HR		Total
	Germany	Italy	Germany	Italy	
None	114	111	119	116	460
Fair	117	122	116	119	474
Discriminatory	124	117	118	118	477
Total	355	350	353	353	1411

Source: Own analysis

We now explain the chronology of the decider experiment, as represented in Figure 5.

Figure 5 Chronology of the deciders' experiment.



Source: Own elaboration

**Explanation of the task:** Participants were told they would be shown a succession of 12 pairs of applicants and asked for each of those pair which of the two applicants they wanted to hire/lend to. The dimensions to judge applicants coincide with the variables used by the AI-based DSS as inputs for decision, although the number of categories may be smaller than those provided to the AI. This is to align with explanations given via LIME, which sometime put different categories (e.g. for age) together. The variables and categories were:

1. Gender: Male or Female,
2. Age: 18-34, 35-54, 55-65,
3. Nationality: German or Italian,
4. Level of education: Low (up to high school), Middle (up to a bachelor) or High (masters and more),
5. Income: Low (less than median income), Middle, High (top 20%), Unknown,
6. Interview score (Bad, OK, Good, Very Good).<sup>9</sup>

<sup>9</sup> Interview score was presented as based on answers to a questionnaire that evaluated a participant's degree of motivation and self-confidence, and was based on questions Q2\_1 to Q2\_5 in the recipient questionnaire (See Annex 2).

Incentives: Deciders in HR were explained the real effort (summing) task done by recipients and told they would get 4 points for each sum made correctly by the person they hired, and the person they hired also got a wage of 100 points (4.3 euro) from them. Their net payoff would therefore be (correct sums \* 4 points) – (100 points). They were made aware that the higher the score of the person they hired, the more they would earn and told that job applicants manage 64 correct sums on average, but this can vary between 34 and 107 depending on the job applicant. A table showed them their payoff in points and euros depending on the number of correct sums. The rate was 100 points = 4.3 euros.

Deciders in banking were explained the trust game in the same way as it was explained to recipients, and told the amount repaid by recipients was 120 points on average, but this could vary between 50 and 150 points depending on the loan applicant. A table showed them their payoff in those different situations. The amount paid back was converted into euros at the rate of 100 points = 4.3 euros and paid to deciders.

Participants had to answer 4 questions to check their understanding.

Preference elicitation: We then elicited preferences of the decision makers among candidates prior to asking them to make choices. They were asked, for each dimension, if that dimension was of High, Moderate or Low importance for them, or Irrelevant. For each dimension they rated as not irrelevant, they were then asked which type of applicant they favoured most (as per the categories presented above).

We then recapitulated participants' preferences with a table. We show an example in Table 3. This was shown to them, with importance colour coded with darker shades indicating more importance.<sup>10</sup> This was also recapitulated in writing.

Table 3 Presentation of decider's own preferences, example.

Variables	Importance	Preferred type
Gender	High	Male
Age	Middle	[35-54]
Nationality	Low	German
Level of education	High	Middle
Income	Irrelevant	
Interview	Middle	Very Good

Source: Own material

The AI-based DSS: For participants who got support from an AI DSS, we then explained its programming and preferences. We told them the DSS predicted the performance of job applicants in the summing task / trust game based on their personal characteristics. We told them this included protected characteristics, such as gender or nationality.

---

<sup>10</sup> We offered participants two alternative choices of colour gradations depending on their preference, one of them designed to address colour blindness.

- In the case of the discriminatory DSS, we told them the DSS was programmed to INCLUDE THE IMPACT of those variables in a job applicant's grade, and that relying on the DSS may therefore lead them to discriminate across job applicants based on protected characteristics.
- In the case of the fair DSS, we told them the DSS was programmed to MINIMIZE THE IMPACT of those variables on a job applicant's grade, and that relying on the DSS therefore ensures they do NOT discriminate across job applicants based on protected characteristics.

They were told they were free to choose according to the grade given by the DSS, or not.

Preferences of the AI were presented to deciders in the same format as their own preferences were recapitulated to them, cf. Table 3. Annex 3 outlines the preferences of the four different AIs (one fair and one discriminatory for both sectors (HR/Banking). The main difference between fair and discriminatory AI was that gender and nationality were rated as irrelevant by the fair AI, and rated as relevant, with different degrees of importance, by the discriminatory AI.

Decision making: We finally explained how recommendations would be presented and decisions elicited. We show an example in Table 4.

Table 4 The decision interface showing explanations for the AI recommendation, example.

	Job applicant A	Job applicant B	
ID	115		269
Gender	Male	+	Female
Age	18-34	+	=
Nationality	German	+	German
Level of Education	Middle	=	Middle
Income	High	++	High
Interview	Low	-	Low
Overall grade	+	-	
	I want to hire A <input type="checkbox"/>	I want to hire B <input type="checkbox"/>	

Source: Own material

Following those explanations and a few comprehension checks, participants were shown a series of 12 pairs of applicants, and their choices were recorded. Applicants were paired to obtain pairs that differed in one of the protected characteristics (e.g. male vs female, but of the same country, or Italian vs. German, but of the same gender), with some exceptions, and that differed in overall grade by no more than two levels (e.g. someone rated ++ with someone rated =, but not someone rated ++ with someone rated -).

Questionnaire: After those choices were made, participants were asked a series of questions about their decisions, attitudes, and background (see Annex 2). In addition to age, gender, country,

region, education, nationality, occupation, sector of employment, household monthly income and social class, we asked them how long they worked in HRM/banking, their hierarchical position (number of employees reporting to them) and size of company (Q15 to Q17), reliance on data and DSS in their job (Q18 to Q20), diversity and diversity policies in their company (Q21 to Q23).

We also asked participants questions about their goals, priorities, and confidence when doing the task (Q3 to Q6), their perceptions of the compared honesty, work ethic, reliability and performance of men and women, and of Italians and Germans (Q7 to Q8), their view on discriminating by gender or nationality (Q9 to Q10), and their perception of the DSS (Q11 to Q14).

At the end of the experiment, participants were asked if they would be interested in taking part in further activities related to the experiment, where they would take part in discussions with other participants and the researchers who designed the experiment and speak about their experience in the experiment and about the use of artificial intelligence in human resource management/banking. They were told they would be compensated for the time spent participating in these activities and given a link to a survey hosted by the European Commission where they could give their email address. We used the pool of volunteers from this stage for the qualitative interviews that followed the experiment.

#### 4.1.4 Payment of incentives

For this study, besides the regular participation fee paid by IPSOS, respondents received points based on their performance in the survey and on the decisions of other respondents. The additional points received were converted at a rate of 100 points = 4.3 euros.

**Recipients:** Incentives were paid out to recipients after we ran the decider side of the experiment. We paid each recipient based on the decisions made by a randomly drawn sample of deciders from both HR and Banking (equal in size to the number of recipients). Recipients were paid in proportion to how many deciders hired or lent to them, and based on what they said they would repay of their loan. As mentioned above, points were converted into euros at a rate of 100 points=4.3 euros.

Out of the 528 recipients that completed the survey, 316 were hired/lent money to, and received an additional performance-based incentive on top of the participation fee. The other 212 recipients were not hired/lent money to and received only the participation fee (paid by Ipsos in points redeemable in their own online shop). On average, recipients received 5.8€ per person (not including the monetary equivalent of the participation fee paid by Ipsos).

**Deciders:** We paid each decider based on the score (HR) or money repaid (banking) by the recipient they selected in one randomly drawn of the 12 hiring/lending decision they made (see "incentives" in the chronology of the experiment). For the deciders' hiring experiment, 658 respondents received an additional performance-based incentive, while 48 only received the standard participation fee because their chosen applicant repaid nothing of their loan. On average, HR participants received 6.0€ per person. Similarly, in the deciders' lending experiment, 645 respondents received an extra incentive, while 60 received the standard participation fee because their chosen applicant did less than 25 sums correct. On average, banking participants received 4.4€.

## 4.2 Post-experimental qualitative studies

We followed up the experimental study with qualitative studies. A sub-set of the experiment's participants (N=13) were invited to analyse the experiment and their experience in individual interviews and groups sessions. Following the participants' second round of engagement, we ran a speculative workshop with a multidisciplinary group of researchers and actors in the AI fairness debate (N=14). The aims of the qualitative studies were threefold:

1. First, we aimed to unpack participants' behaviours in the context of specific scenarios of AI-supported decision-making relevant to their profession. This would provide contextual information regarding the demographics of the specific participants, the level of their digital literacy and their current practices of the use of AI for personal and professional goals as well as to give the context and the space for any additional indirect insights, participants wanted to share with us. In addition, during the interviews, participants had the opportunity to interact with a biased and unbiased AI system. After the interaction, we prompted them to reflect on the recommendations they received. Through the conversation, we aimed to elicit possible assumptions they make and to explore the level of consciousness about their personal or other biases.
2. Second, we aimed to identify elements that would inform the ecological validity of this study, meaning the level of generalization of this study to participants' real-life professional contexts. Although the term ecological validity has specific use in the field of psychology such as the examination of the suitability of the research tools (Schmuckler 2001), for this research we take a more general approach and use the term to communicate the relevance of the specific study and scenarios with real-life situations in the context of participants' professional settings.
3. Third, in addition to the unpacking of participants' current behaviours and attitudes toward AI-based decision-making support systems, we were interested in understanding how specialists envision these systems to develop in the future in a way that would promote fairness and overcome possible biases, not only from AI but also from humans, institutions and our society in the specific scenarios of human resources and banking but also beyond the specific scenarios.

Those qualitative studies aim to gain in-depth insight into the impact of human oversight in situations in which professionals use AI-based systems as a supportive tool for their decision making.

More specifically, we investigated the following Research Questions (RQs):

- RQ1: Are people willing to use AI-support for their decision-making in the specific scenarios and for what reasons?
- RQ2: Are participants aware of their own and algorithmic biases and what kind of measures do they take to mitigate these biases?
- RQ3: What kind of contextual factors affect people's decision-making process in their real-life scenarios?

In addition, the study aims to identify future directions for the design of Decision Support Systems that would contribute to fairer decision-making. As such, we aimed to answer the following additional RQ:

- RQ4: How can we envision a fairer hybrid system of algorithm-supported human decision-making process in the scenarios under examination and in other real-life scenarios?

#### 4.2.1 Individual interviews

The first part of the qualitative study focused on the interviews and the workshops with a subset of the professional participants.

Based on the research questions, we designed the structure and the corresponding questions for the semi-structured interviews. Semi-structured interviews enable reciprocity between the interviewer and participant and allow space for participants' narratives on the topic of research (Smith 1995). In semi-structured interviews, questions function as an opportunity for each participant to provide certain input and according to common practices, the interviewer adapts the questions to optimize the flow of the conversation, while participants are encouraged to provide additional relevant information if they want to.

Initially, we aimed for a first round of total number of sixteen interviews, to cover the four (2X2) basic demographics of our study, country (Italy and Germany) and profession (Human Resources and Banking). This was decided for us to situate and construct diversity in the pool of participants according to the main variables of the study (McIntosh and Morse 2015).

According to the research design, we would examine for data saturation in the first round of interviews (Vasileiou et al. 2018) and decide whether more interviews were needed depending on this. From a methodological point of view, in interviews, data saturation is reached when the research team notices the same or similar themes appearing repeatedly during the conduction of interviews or the data analysis and no new themes, ideas, opinions, or patterns appear anymore (Saunders et al. 2018). As such, in the end of the first round of interviews, the research team proceeded with transcriptions and annotation of the data from N=13 interviews.

The data annotation of the first round of interviews indicated data saturation. To reach this conclusion we involved observation of the codes in successive transcripts of the interviews noticing that new code frequency was diminished which signalled the reach of saturation. As such, the research team decided that for the purposes of this study, no further interviews were needed.

Based on the research questions, the interviews were designed to cover three main topics, each being the subject of a different phase in the interview process.

#### 4.2.2 Phase 1: Contextual information and current practices on the use of AI

Contextual information is considered to play a central role since it can be used for the possible identification of interconnections and would be useful for the interpretation of the results of the study (Roller and Lavrakas 2015). In addition, for the interpretation of participants' attitudes towards the use of AI, it was important for us to understand their current practices and predispositions. As such, we asked specific questions to extract contextual information and information about participants' current practices on the use of AI (Table 5). The questions focused on:

- the collection of additional demographics (questions 1-6),
- the current practices of the participants in relation to the use of AI at work (questions 7-13) and
- the possible policies on AI ethics promoted by participants' companies (question 14).

Table 5 Phase 1 of the interview: Questions that contribute to the identification of contextual information.

Questions for contextual information and current practices on the use of AI
<ol style="list-style-type: none"> <li>1. Where do you live?</li> <li>2. Where do you work?</li> <li>3. How large is your organization?</li> <li>4. What is your role in the company?</li> <li>5. For how long have you been working there?</li> <li>6. What is the conformation of the company?</li> </ol> <ol style="list-style-type: none"> <li>7. Which devices do you use in your everyday life?</li> <li>8. What do you use them for?</li> <li>9. How do you use them?</li> <li>10. Which products are you aware that use AI?</li> <li>11. Are you using any of them?</li> <li>12. Do you use any AI tools or plug-ins at work? For examples: a tool that process CVs/Creditworthiness. What do you think about them?</li> <li>13. Do you do programming? How comfortable are you with data at work?</li> </ol> <ol style="list-style-type: none"> <li>14. Any training or policy on AI, ethics, discriminations etc?</li> </ol>

#### 4.2.3 Phase 2: Interaction with a biased/unbiased system

In this phase, the participants interacted with a simulation of the interface and the activity of the experimental study. As shown in Table 6, first the participants were asked to recall their experience during the experimental study, then they were asked to repeat an instance of the experimental study by interacting with the simulation of the AI-supported decision-making system to select a candidate (adapted to the profession of the participant) based on certain characteristics of the candidate. Since the main goal of this activity was to understand the rationale of the participants behind their decision-making, we asked the participants to think out loud during the process. To support their reflection during the interaction, we prompt the participants with certain question (questions 2a and 2b).

Table 6 Phase 2 of the interview: Questions to support the participants' reflection about their interaction with the algorithm.

Interaction with biased/unbiased system
<ol style="list-style-type: none"> <li>1. Recall from the behaviour experiment.             <ol style="list-style-type: none"> <li>a. Do you remember what it was about?</li> <li>b. Can you describe what you had to do during the experiment?</li> </ol> </li> <li>2. Interaction with AI             <ol style="list-style-type: none"> <li>a. Here's an example of 2 profiles of candidates. What do you see here? Which assumptions can you make about these people?</li> <li>b. (HR) Which one do you think it would perform best for a mathematical task? / (FIN) Which one would you trust more for repaying a loan?</li> </ol> </li> <li>3. Reveal AI             <ol style="list-style-type: none"> <li>a. What do you see? What is this information telling you?</li> <li>b. What do think about this?</li> <li>c. What is the recommendation here?</li> <li>d. Would you follow this recommendation?</li> <li>e. Why?</li> <li>f. Do you think this suggestion is fair?</li> </ol> </li> <li>4. Explanation about the training of the system             <ol style="list-style-type: none"> <li>a. Had you noticed anything weird or suspicious? Why?</li> <li>b. What do you think? Do you think the grading is fair?</li> <li>c. Would you change your opinion?</li> </ol> </li> </ol>

After the selection, we revealed the algorithm used by the AI to support certain participants (Figure 6). The algorithm (similarly to the experiment) could be biased or unbiased. We prompted the participants' reflection on the revealing of the algorithm with questions 3a – 3f as listed in Table 6.

Figure 6 Interface used for the examination of algorithmic biased during the interviews, German version.

SCENARIO AB	Kandidat	A		B		
Geschlecht	Frau	--		Mann	++	Sehr negativ
Alter	18-34 Jahre alt	=		35-54 Jahre alt	+	Negativ
Nationalität	Deutsch	=		Italienisch	=	Umwichtig
Bildungsniveau	Mittel	+		Hoch	+	positiv
Einkommen	Hoch	++		Mittel	-	Sehr positiv
Interview	Großartig	+		Gut	=	
Gesamt		=			+	

Source: Own material

Following the revealing of the algorithm, we explained to the participants the way we trained the system and we sought their opinion and critical reflection with the supporting questions 4a – 4c.

#### 4.2.4 Phase 3: Reflection on priorities and biases in decision making in the specific scenarios

With the last phase of the interview, we sought to understand how the interaction with the biased or unbiased algorithm and the explanations we gave the participants shaped their opinion about the use of AI in his professional environment and their sense of control and oversight when interacting with an algorithm (Table 7).

Table 7 Phase 3 of interview: Questions to support participant reflection about the explanations provided during the activity in phase 2.

Priorities and biases for picking candidates from suggestions
a. Do these explanations make sense to you?
b. What do you think about both AI systems? Which one do you think is the fairer? Which one would you prefer or trust more?
c. What do you think about what you see? Can you think out loud?
d. Are the characteristics that are important for the AI also important for you?
e. Which one would you prefer using? Why?

#### 4.2.5 Group interviews

Following the interviews, the same participants were invited to participate to interactive workshops on the collaborative software Miro, in groups, with discussion via the video communication platform (Figure 7).

The discussions aimed at providing the participants with the opportunity to reflect on the experiment and the interviews and specially to explore the ecological validity of the study in the context of their professional activities. The session was also an occasion to explore in-depth their assumptions and participants' awareness about their own assumptions and biases.

Figure 7 Screenshot from the videocall, sharing the screen with the Miro board activity for the workshop with participants.



Source: Own material

For this interactive workshop, the participants took part to two main activities. The first one was on analysing and commenting the preliminary results of the quantitative study. The goal of the activity was to gather their point of view of what might have led to those results. The main facilitator of the session presented the results as a form of quiz, using several statements from the study and asking participants to vote a possible outcome. This exercise served to both investigate the results with the relevant stakeholders as well as gather more evidence from the other aspects people listed while interpreting in an engaging and participatory way where all the participants intervene.

For the second part, the activity focused on ecological validity of the experiment, asking participant to review the online experiment through screenshots while describing and comparing them to the real-life situations. This activity helped us to gather feedback on the specific elements of the experiments as well as contextual factors related to people's jobs and approaches and how they would have seen the experiment play out in their daily lives.

#### 4.2.6 Experts workshop

As the last step of our qualitative part of the study, we conducted an ideation co-design workshop with a multidisciplinary team of experts. After exploring the ‘whys’ with the participants, we wanted to bring the discussion and insights into context, specifically looking at implications for the future, and for the policy initiatives of the European Commission. In order to look at the futures, and its consequences, we selected a multidisciplinary group of researchers and actors in the AI fairness debate to discuss the results of our experiment. We selected this group and designed a workshop in line with the co-design participatory practice for its adequate approach in creating a format through which diverse stakeholders can share their ideas, become exposed to the ideas of others, and generate new ideas. As a methodology, co-design is more active and hands-on than other methodologies that are common in public policy (Forlano & Mathew, 2014, pages 2-3).

This study differs from traditional social science methods as it uses a participatory design approach to structure part of the engagement with participants as well as experts. Research methods in this case includes a mix of written, visual, verbal, and observational strategies. Within Participatory Research, people might be involved at any given steps or throughout the whole process. Activities might include templates, tools, and specific tasks, and activities to enable participation, and shared decision-making. These research methods, as a broadly agreed definition, include tools and techniques used throughout the process of data collection, as well as data analysis, interpretation and dissemination (Vaughn and Jacquez 2020). The research team was motivated by experimenting with a mix of social science and design methodologies. For this reason, when focusing on the future focused side of the activity, the team decided to tap into speculative design and design fiction, which reach beyond identifying needs and solving problems, but rather, move towards a generative and future-oriented space of alternative options (Forlano & Mathew, 2014, page 10) We designed this co-design participatory workshop through the lenses of the speculative design practice, a type of design that allows designers to imagine and explore different possible futures or alternative realities (Auger 2013). By examining the ways in which technology, culture, society, and other factors interact and influence one another, we gathered to challenge our current assumptions about the present and help us to anticipate and prepare for the future.

The 1-day workshop took place in November at the collaborative space of the EU Policy Lab, in the Joined Research Centre’s Building of the European Commission in Brussels (BE). The aim of the workshop was to ideate how to mitigate human and algorithmic biases in hybrid AI-supported decision-making, generating ideas of interventions, with future visions of their implications. During the session with experts, we investigated the contextual needs for a fairer system as well as envisioning interventions. Through this workshop the research team wanted to discover and understand how AI-supported human decision-making in the specific predefined scenarios of hiring and loaning and beyond, is perceived by experts of different scientific disciplines. Often certain disciplines come with certain epistemological paradigms that function as a lens for an expert to explore a specific topic, choose certain methodologies for data collection and analysis, and interpret the data they collect or, in more theoretical disciplines, conceptualize, reason, and analyse a certain concept. As many studies have noted, the results generated by the likely frictions and tensions that might arise from a participatory design process, when diverse groups are convened, may be a valuable aspect of the interaction (Tsing 2004). With this workshop, the research team not only aimed to capture the perspectives of each expert but also to create opportunities for all the experts as a group to interact with each other in various settings and create an amalgam of opinions crafted through their interactions.

The issues this workshop raised prompted us to analyse and discuss institutional decision making, and fairness as a moral lens to analyse these situations. Our study focused in specific on the concrete examples of recruitment and loans, typical examples of potential loss of opportunities caused in scenarios where someone is taking a decision that might result in providing or not opportunities to others. Alongside these, in this workshop, we included other two scenarios (education, and public surveillance), and asked experts to make considerations on risks levels. The

day of the workshop was structured around reflections on human oversight in current practices and future aspirations. Table 8 shows the outline of the workshop.

Table 8 Structure of the 1-day workshop with Experts

Introduction
Presentation of the Quantitative study and discussion of current practices
<b>Defining the notion of “Fairness”</b>
Future aspirations on Human-AI interaction for decision-making
Close

Figure 8 Workshop with experts during the presentation of the outcome of activity 1



Source: Own material

During the day, the first activity we ran was about envisioning fairness, specifically in the context of Human-AI collaborative decision making. We asked participants to design some 'tarot cards' in small groups, symbolizing the past, present, and future of this concept, and to present the results to the group (Figure 9). Each card needed to include metaphors using existing theories, personal work, and reflections. The aim of this activity was to work on a metaphorical as well as visual level to encourage the participants to ground their ideas and concepts in examples.

We then continued the day moving to short-term future scenarios, providing tools to encourage them to analyse the cases and design possible future interventions. We created 4 groups and scenarios, based on the topics of job searching, loans giving (Figure 9), education and surveillance. Each group prepared the overview of what this intervention could have looked like for a hypothetical "Fair AI + Human Catalogue". Each team worked on defining the key characteristics of possible interventions and considerations for these future scenarios, including stakeholders, societal norms, interaction, algorithm design, and the policy context.

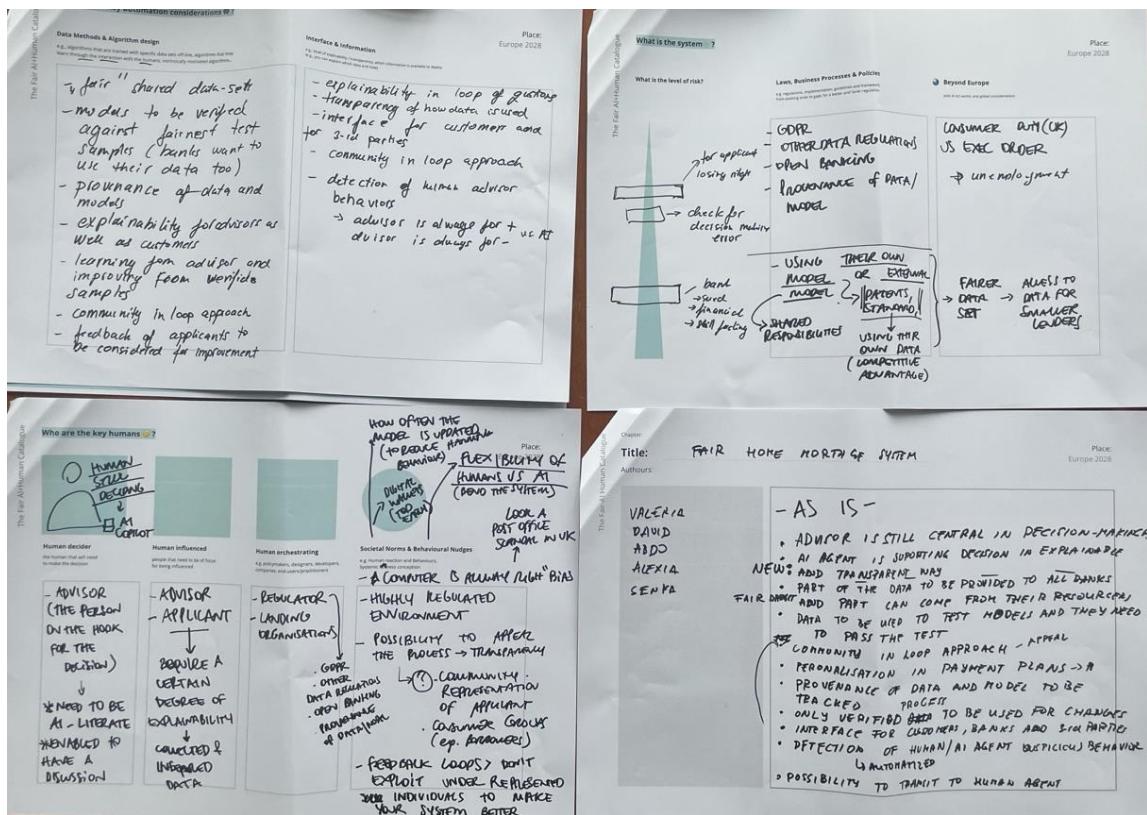
The day concluded with the different experts picking the key priorities for further discussion and analysis at the commission. We asked the group to make some reflections, this time divided by same type of professions/disciplines, and to list what ideally, they would have like to see more of

from the work of the Commission on the topic. Each group drafted a 'mission letter' with their recommendations, suggesting reviews, analysis, and possible interventions.

The data collected during the workshop were in the form of paper material, participants' notes, sketches, sticky notes etc. All materials collected after the end of the workshop were digitalized by members of the research team shortly after. Members of the research team discussed, organised, and categorised the data for further elaboration.

It should be noted that since the material was created collectively, in the following analysis of the data we are not able to identify the exact person that contributed with a specific quotation. In addition, often a thought written in the provided forms was a product of extensive discussion among the participants and as such they are not allocated to a single person.

Figure 9 Workshop with experts, output of the tools for activity 2 for the mortgage scenario



Source: Own material

## 5 Results of the experiment (Quantitative)

We first describe the sample of deciders, their socio-demographic characteristics, work environment and experience, perception of the experiment, preferences among candidates, and their level of prejudice by country and gender. We will then analyse their choices during the experiment, and consider their level of reliance on AI, how AIs affect their decisions, and how AI and decider preferences interact.

### 5.1 Descriptive statistics

#### 5.1.1 Socio-demographics

By design, gender was balanced with an equal representation of male and female deciders in our sample (Table 9). Most participants (about 70%) were in the 35-54 age group. There were no significant differences in age distribution by sector, but deciders in HR in Italy were younger than those in Germany. Education levels were significantly higher in Germany than in Italy ( $p$  value < 0.0001, Pearson's Chi-squared test), and this in both sectors.<sup>11</sup> Income categories were defined with respect to the overall income distribution in each country<sup>12</sup>. About 40% of deciders have “high” income (top quintile in their country), and about 35% have “middle” income (above the median but less than the top quintile). Italian deciders are significantly more likely than German deciders to belong to the top income category. Differences between sectors are not significant.

Table 9 Socio-demographics, deciders

		Banking		HR	
		Germany	Italy	Germany	Italy
Gender	Male	184 (52%)	168 (48%)	168 (48%)	178 (50%)
	Female	171 (48%)	182 (52%)	185 (52%)	175 (50%)
Age	18-34	98 (28%)	81 (23%)	69 (20%)	109 (31%)
	35-54	251 (71%)	250 (71%)	268 (76%)	227 (64%)
	55-65	6 (2%)	19 (5%)	16 (5%)	17 (5%)
Education	Low	5 (1%)	26 (7%)	5 (1%)	40 (11%)
	Middle	62 (17%)	174 (50%)	68 (19%)	178 (50%)
	High	288 (81%)	150 (43%)	280 (79%)	135 (38%)
Income	Low	102 (29%)	65 (19%)	89 (25%)	99 (28%)
	Middle	144 (41%)	110 (31%)	129 (37%)	104 (29%)
	High	108 (30%)	168 (48%)	135 (38%)	145 (41%)
	Unknown	1 (0.3%)	7 (2%)	0 (0%)	5 (1%)

Source: Own analysis

<sup>11</sup> Education levels are defined as Low (up to high school), Middle (up to a bachelor) or High (masters and more).

<sup>12</sup> Low income is defined as less than 2200€ in Germany and less than 1650€ in Italy. This is less than the median income in the country, in the same way as low recipient income was defined for recipients. High income is defined as more than 2850€ in Germany and more than 2200€ in Italy. This is income in the top 20% in the country, in the same way as high recipient income was defined for recipients.

### 5.1.2 Work environment.

Considering now the work environment (Table 10), we find that deciders work in relatively large companies (42% with more than 250 employees, 46% between 50 and 249 employees). They are also experienced (44% have more than 5 years' experience) and are in higher hierarchical levels (19% with more than 20 employees reporting to them, only 5% with no one reporting to them). We do not notice large differences in those statistics across sectors. However, we do observe differences across countries, whereby Italian participants are in higher hierarchical levels on average, in terms of number of people reporting to them) and work in firms of bigger size (in terms of number of employees).

Table 10 Work environment and experience.

		Banking		HR	
		Germany	Italy	Germany	Italy
Work experience	Less than one year	11 (3%)	13 (4%)	18 (5%)	18 (5%)
	Between one and two years	72 (20%)	59 (17%)	58 (16%)	47 (13%)
	Between three and five years	132 (37%)	120 (34%)	124 (35%)	119 (34%)
	More than five years	140 (39%)	158 (45%)	153 (43%)	169 (48%)
Hierarchical level	No one reports to me	17 (5%)	28 (8%)	8 (2%)	23 (7%)
	Between 1 and 5 employees	65 (18%)	45 (13%)	46 (13%)	41 (12%)
	Between 6 and 10 employees	139 (39%)	104 (30%)	117 (33%)	91 (26%)
	Between 11 and 20 employees	95 (27%)	106 (30%)	100 (28%)	113 (32%)
	More than 20 employees	39 (11%)	66 (19%)	82 (23%)	82 (23%)
	I don't know	0 (0%)	1 (0.3%)	0 (0%)	3 (1%)
Company size	Less than 10 employees	4 (1%)	2 (1%)	9 (3%)	13 (4%)
	10-49 employees	48 (14%)	41 (12%)	34 (10%)	21 (6%)
	50-249 employees	184 (52%)	136 (39%)	155 (44%)	169 (48%)
	More than 250 employees	118 (33%)	169 (48%)	155 (44%)	147 (42%)
	I don't know	1 (0.3%)	2 (1%)	0 (0%)	3 (1%)

Source: Own analysis

Participants are also experienced in dealing with data and statistics in their job (Table 11). We find that 45% deal very often with them, 47% only sometimes. Many have experience using decision support systems in their organisation (15% use them very often, 40% use them often). We find that Italian participants deal more often with data and statistics and are less likely never to have used a decision support system in their work.

Table 11 Experience with data and DSS

		Banking		HR	
		Germany	Italy	Germany	Italy
Data experience	Very often	146 (41%)	171 (49%)	156 (44%)	165 (47%)
	Sometimes	186 (52%)	145 (41%)	171 (48%)	156 (44%)
	Rarely	11 (3%)	22 (6%)	20 (6%)	22 (6%)
	Never	12 (3%)	12 (3%)	6 (2%)	10 (3%)
DSS experience	Very often	48 (14%)	44 (13%)	64 (18%)	53 (15%)
	Sometimes	132 (37%)	156 (45%)	124 (35%)	156 (44%)
	Rarely	45 (13%)	63 (18%)	69 (20%)	75 (21%)
	Never	115 (32%)	26 (7%)	87 (25%)	49 (14%)
	I don't know	15 (4%)	61 (17%)	9 (3%)	20 (6%)

Source: Own analysis

In terms of diversity, 80% of participants state there is at least “some” diversity in their company (Table 12), and 71% state there is a diversity policy in their company. A total of 77% judge that this policy is well or very well implemented. We find some country differences, whereby participants in Germany are more likely to think there is a lot of diversity in their company, report more often that there is a diversity policy, and that it is well or very well implemented.

Table 12 Company diversity

		Banking		HR	
		Germany	Italy	Germany	Italy
Diversity in company	Yes, there is a lot of diversity	150 (42%)	97 (28%)	133 (38%)	76 (22%)
	Yes, there is some diversity	160 (45%)	163 (47%)	163 (46%)	195 (55%)
	No, there is not much diversity	34 (10%)	58 (17%)	33 (9%)	56 (16%)
	No, there is no diversity	11 (3%)	32 (9%)	24 (7%)	26 (7%)
Diversity policy	Yes	271 (76%)	235 (67%)	267 (76%)	230 (65%)
	No	77 (22%)	82 (23%)	78 (22%)	94 (27%)
	I don't know	7 (2%)	33 (9%)	8 (2%)	29 (8%)
Implementation of diversity policy	Very well	90 (33%)	88 (37%)	123 (46%)	81 (35%)
	Well	139 (51%)	90 (38%)	90 (34%)	77 (33%)
	Average	36 (13%)	45 (19%)	47 (18%)	67 (29%)

	Banking		HR	
	Germany	Italy	Germany	Italy
Badly	4 (1%)	9 (4%)	6 (2%)	4 (2%)
Very badly	0 (0%)	0 (0%)	0 (0%)	1 (0.4%)
I don't know	2 (1%)	3 (1%)	1 (0.4%)	0 (0%)
NA	84	115	86	123

Source: Own analysis

### 5.1.3 Decisions and reliance on the DSS

We asked participants several questions about the way they made decisions in the experiment (

Table 13). They were invited to say how important fairness was to them compared to efficiency (from 1 to 4, higher means more importance given to fairness), whether they trusted their instinct or rationality, if it was more important to make correct rather than fast decisions, and whether they were confident in their choice. With the value 2.5 indicating equal importance given to both dimensions, we find that efficiency was about as important as fairness for participants (mean of 2.7), and that they relied about as much on reason as on instinct (mean of 2.5). However, they put more emphasis on making correct rather than fast decisions (mean of 3.3). They also were relatively confident in their choice (mean of 3.3). Differences across sectors and country were small, except in terms of the reliance on rationality, which was higher among Italian participants, especially in the HR sector.

Table 13 Mode of decisions

	Banking		HR		
	Germany	Italy	Germany	Italy	
Importance of fairness vs. efficiency	Mean (s.d.)	2.9 (0.9)	2.7 (1.0)	2.7 (0.9)	2.6 (1.0)
	N	355	350	353	353
Reliance on instinct vs. reason	Mean (s.d.)	2.7 (1.0)	2.5 (1.0)	2.7 (0.9)	2.2 (0.9)
	N	355	350	353	353
Goal to be correct vs. fast	Mean (s.d.)	3.3 (0.7)	3.3 (0.8)	3.3 (0.8)	3.4 (0.7)
	N	355	350	353	353
Level of confidence in choice	Mean (s.d.)	3.2 (0.7)	3.3 (0.6)	3.2 (0.7)	3.3 (0.6)
	N	353	349	351	351

Source: Own analysis

In terms of reliance and evaluation of the AI (Table 14), we find that on a scale from 1 indicating "not at all" to 4 indicating "to a large extent", participants were split in their reliance on the AI (mean of 2.7), while they indicated they "somewhat" understood the AI (mean of 3.0). They were also relatively positive regarding whether the AI was fair and accurate. We find also that participants in HR were more likely to rely on the AI, reported better understanding, and rated the AI as fairer and

as more accurate than participants in banking. We did not find significant differences by country, and also no differences between treatments (i.e. whether the AI was fair or discriminatory).

Table 14 Evaluation of the AI

		Banking		HR	
		Germany	Italy	Germany	Italy
I relied on AI	Mean (s.d.)	2.6 (0.9)	2.5 (1.0)	2.9 (0.9)	2.8 (0.8)
	N	241	239	234	237
I understood AI	Mean (s.d.)	2.8 (0.9)	2.8 (1.0)	3.1 (0.8)	3.2 (0.8)
	N	241	239	234	237
AI was fair	Mean (s.d.)	2.7 (0.9)	2.7 (1.0)	3.1 (0.9)	3.0 (0.8)
	N	241	239	234	237
AI was accurate	Mean (s.d.)	2.7 (0.9)	2.8 (1.0)	3.0 (0.8)	3.1 (0.8)
	N	241	239	234	237

Source: Own analysis

#### 5.1.4 Discriminatory preferences

We elicited deciders' discriminatory preferences by asking them how important each candidate characteristics were to them, and which type of candidate they favoured. This corresponds to how preferences of the AI are subsequently presented to them.

Deciders rated interview, income, and education as equally important (mean of 2.4), followed by age and nationality (mean of 2), with gender the least important (mean of 1.7). Only 17% of deciders rated gender as irrelevant, however, 10% for nationality, 6% for age.

51% of deciders prefer male applicants, 60% prefer Germans, 63% prefer applicants aged between 32 and 46 (only 1% prefer participants older than 46), 63% prefer highly educated participants (only 2% prefer participants with low education), and 57% prefer participants with high incomes (only 2% prefer those with low incomes). Deciders split between preferring participants with good or very good interview scores.

There are limited differences between sectors in terms of preferences, but deciders in banking gave less importance to education and interview and more importance to income. The preference for male and young applicants was pronounced in human resources, while female and middle-aged applicants were on average preferred in banking, where the preference for high income applicants was also higher.

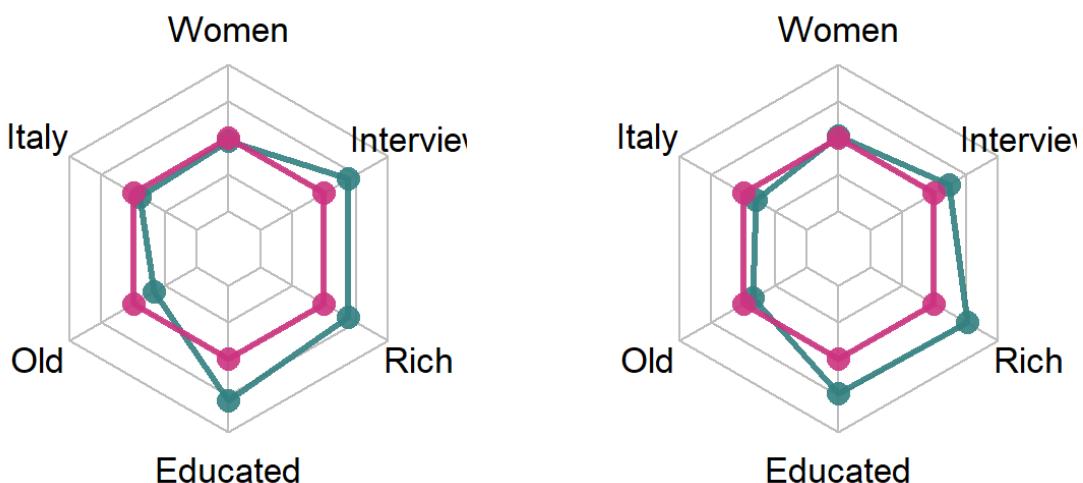
We summarize deciders' preferences by multiplying the importance of a dimension, from Irrelevant, Low, Moderate, High, graded from 0 to 3, and the direction of the preference, from -1 to 1, indicating the extremities of the characteristics. For example, if asked to choose between Men and Women, -1 codes a preference for men, 1 a preference for women. If asked to state a preference in terms of the age of the applicant, 18-34 year old are coded as -1, 35-54 year old as 0 and 55-65 year old as 1. The preference measure can thus range from -3 to 3 for each dimension.

Figure 10 represents this preference measure for applicant characteristics by country and sector. We see a clear home bias, whereby Germans prefer German applicants, and Italians prefer Italian applicants.

Figure 10 Deciders' preferences among applicants, by sector.

**HR**

**Banking**



*Source: Own analysis*

The preferences of the deciders are broadly consistent with those of the fair AIs on average in terms of gender, country, age, or interview score (cf. Figure 22 in Annex 3). Deciders have however on average a higher preference for high levels of education and wealth. The preferences of the discriminatory AI are consistent with the average preferences of male German deciders, who also on average prefer male and German applicants (cf. Figure 23 in Annex 4).

Figure 24 in Annex 4 outlines evidence of homophily, whereby deciders prefer applicants similar to them in terms of nationality, gender, and age. However, this does not extend to preferences in terms of education and income.

Gender and nationality prejudice was elicited by asking which of either nationality or gender was more honest, had better work ethic, was more reliable and performed better. Table 15 shows the mean of those evaluations, whereby prejudice vs. women ranges from 1 to 3, with 1 indicating women rate better and 3 indicating men rate better. We find that deciders do not hold on average prejudice for or against women and Italians, whereby values of prejudice are close to 2, meaning both genders and nationality rate equally. However, deciders in banking, and deciders in Germany are slightly (but significantly) more prejudiced against women and Italians. Levels of prejudice against women are similar in men and women.

Table 15 Prejudice and views on discrimination

		Banking		HR	
		Germany	Italy	Germany	Italy
Prejudice vs. Women	Mean (s.d.)	2.2 (0.4)	2.0 (0.4)	2.1 (0.3)	2.0 (0.4)
	N	355	350	353	353
Prejudice vs. Italians	Mean (s.d.)	2.3 (0.3)	2.1 (0.5)	2.2 (0.3)	2.0 (0.5)
	N	355	350	353	353
Is gender Discrimination OK?	Mean (s.d.)	2.2 (0.9)	2.1 (1.0)	2.2 (1.0)	2.1 (1.0)
	N	355	350	353	353
Is country Discrimination OK?	Mean (s.d.)	2.3 (0.9)	2.3 (1.1)	2.4 (1.0)	2.1 (1.0)
	N	355	350	353	353

Source: Own analysis

Finally, when asked whether it was OK to choose applicants based on gender or nationality, on a scale from 1 (never) to 4 (always) deciders reported on average that this was “rarely” the case (Table 15).

## 5.2 Analysis of choices

We proceed with our plan of analysis as pre-registered on the OSF registries at <https://osf.io/5mz3s>. The data and analytic code are available at <https://osf.io/mhd7r>. Our analysis will concentrate on analysing the impact of AI support on discrimination.

However, in order to satiate some readers’ interests, and because previous studies have focused on the impact of AI on performance, not discrimination, we first give some statistics on the performance of deciders compared to the performance of the AI. We compare how many points the deciders earned, on average, through their choice of applicants, and how much they would have earned if they had followed the recommendations of the AI (Table 16). Remember that HR deciders had to hire applicants who perform well in a summing task, and Banking deciders had to lend to applicants who would repay the most out of a loan made to them (cf. section 4.1.3).

Table 16 Deciders' performance vs. AI performance

		Banking			HR		
		Generic AI	Fair AI	No AI	Generic AI	Fair AI	No AI
Points, chosen applicant	mean	109.8	107.2	109.1	59.9	59.9	59.1
	sd	12.2	13.6	13.0	5.0	5.3	4.2
	N	241	239	225	236	235	235
Points, recommended applicant	mean	119.8	114.9	NA	64.8	61.2	NA
	sd	10.3	12.4	NA	4.0	4.1	NA
	N	241	239	0	236	235	0

Source: Own analysis

We find that in the treatment with no AI, banking deciders would have gotten 109 points repaid back out of their 100 points loan, and HR deciders chose applicants who managed to do 59 sums. This is not significantly different from what they would obtain if they had chosen applicants at random (cf. performance and repayments of the average applicant, section 4.1.1).

Furthermore, the performance in selection did not improve when using an AI. On the other hand, we see that following the AI recommendation would have earned them 120 points on average with the generic AI and 115 points with the fair AI. Similarly, they would have selected applicants making 65 sums with a generic AI on average, and 61 points with a fair AI. The potential improvement of following the AI's recommendation is statistically significant in both sectors and for both types of AI.

This finding replicates the now usual finding in the literature, whereby giving access to AI recommendations does not improve human decisions, and humans do worse than if they followed the AI's recommendations all the time (Patrick Hemmer et al. 2021).

### 5.2.1 Choices among applicants without AI support

We now go on properly to focus on the impact of AI on discrimination. We first test whether deciders make discriminatory decisions in terms of gender and country in the absence of AI support. This is to establish our baseline level of discrimination. We thus consider the pattern of choices in the treatment without AI support. We relate the choice of candidate 1 vs. candidate 2 to differences between the characteristics of the two candidates. We report results from a linear probability model<sup>13</sup> for panel data, with random individual effects.<sup>14</sup><sup>15</sup>

---

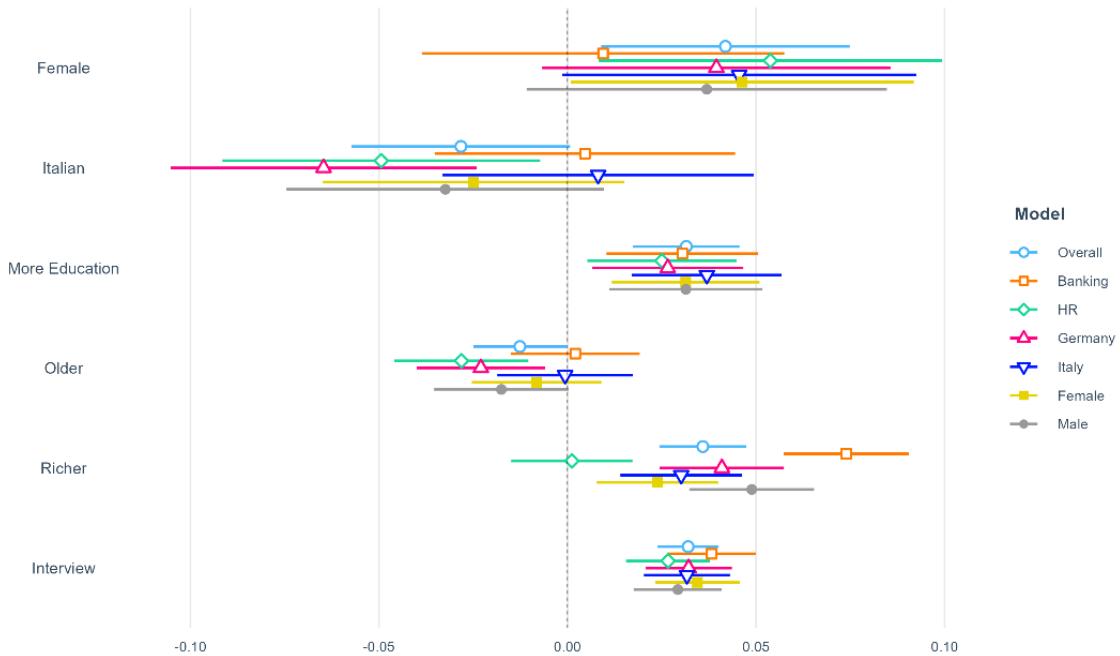
<sup>13</sup> We can use a linear probability model whenever the relationship between probability and log odds is approximately linear over the range of modelled probabilities. In our case, probabilities we investigate are around 50%, which is well within the 20%-80% range where  $\ln(p(1-p))$  is approximately linear.

<sup>14</sup> We tested the assumption of random individual effects with a Hausman-Taylor test.

<sup>15</sup> The estimation equation is of the form

`choice_1=W_vs_M+Ita_vs_Ger+education_diff+age_diff+income_diff+interview_diff` whereby `W_vs_M` is coded a 1 if applicant 1 is a woman and applicant 2 is a man, -1 in the opposite case, 0 else. `Ita_vs_Ger` is computed according to the same principle. Similarly, `education_diff=education_1-education_2`, see Annex 2, "choice variables".

Figure 11 Impact of applicant characteristics on selection, treatment without AI



Source: Own analysis

We find that deciders tend to favour of women and Germans, favour higher levels of education, younger applicants, those with higher income and better interviews. This is consistent with their expressed preferences (see section 5.1.4). There are some exceptions though. Deciders in Banking do not discriminate by gender, country, and age, but put more weight on income. Deciders in HR on the other hand do not put weight on income. Italian deciders do not discriminate by country or by age. Note that both male and female deciders favour female applicants.

We then proceed to test whether deciders are more likely to choose applicants that are more similar to them in terms of their characteristics, and whether they make decisions that are in line with their preferences across applicants, as elicited before the experiment. Regressions shown in Table 24 in Annex 5 thus consider the impact of deciders characteristics and of their expressed preferences on choice.<sup>16</sup> We find that deciders are significantly less likely to choose an applicant from a different country than their own (diff\_diff\_applicant\_country, column 1). Other differences between deciders and applicants do not affect choices significantly. We also consider how closely applicants fit the expressed preferences of deciders and find that deciders indeed favour applicants that more closely correspond to their preferred type and this for all characteristics except gender and education (column 2). This confirms the reliability of the measure of preferences.

Combining deciders characteristics and preferences (column 3), we find they both play a role, which underlines that bias is both conscious (expressed through preferences) and unconscious (working through unacknowledged preferences). In particular, bias against applicants from a different country is driven by one's nationality rather than expressed preferences. Finally, we include expressed prejudice as a factor (column 4), and find prejudice against women does play a role in reducing the share of women chosen (n.s.).

---

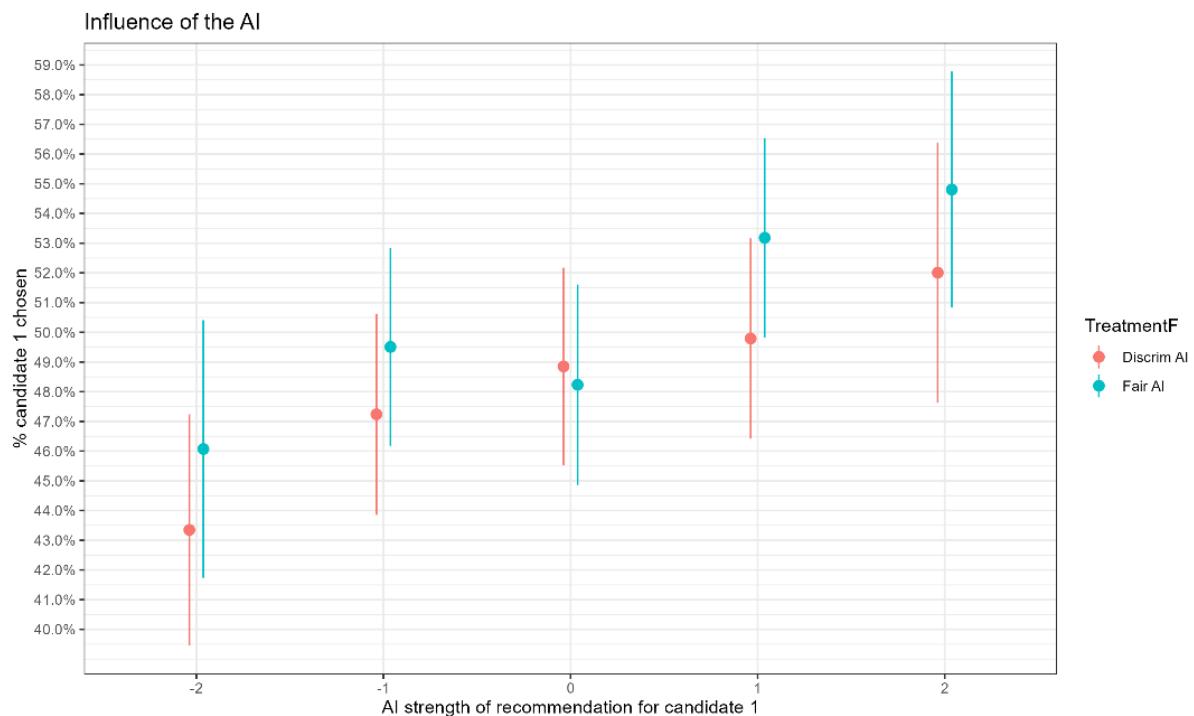
<sup>16</sup> Annex 5 describes variables we use to compute difference between applicant and decider, and between applicant and ideal applicant of the decider. The latter is weighted by how important a dimension is to the deciders.

## 5.2.2 Reliance on AI recommendations

We now test whether deciders are more likely to prefer an applicant if it is recommended by the AI system they are exposed to. We thus look at the extent to which deciders relied on AI recommendations in treatments with a fair and a discriminatory AI. We compare the overall grade given by the AI to applicant 1 and applicant 2 and compute their difference, from -2 to 2, whereby -2 means applicant 1 rated two grades lower than applicant 2. Full compliance with AI recommendation would be such that applicant 1 is never chosen if the difference is less than 0, always chosen if the difference is more than 0, and chosen at a rate close to 50% if the difference is 0.

Figure 12 shows the rate at which applicant 1 was chosen depending on the difference in grades. We find that decision makers follow even the strongest AI recommendations only 55% of the time. They are also not more likely to follow AI that is fair than AI that is discriminatory. This goes against our preregistered hypothesis whereby we expected the fair AI to be followed more often than the generic one. We will consider further this result by linking it to participants' perceptions of the different AI (section 0).

Figure 12 Likelihood to choose a candidate as a function of AI recommendation for or against that candidate.



*Confidence intervals are for the mean of the individual mean % of candidate 1 chosen for each level of the difference in the overall grade given by the AI between candidate 1 and candidate 2.*

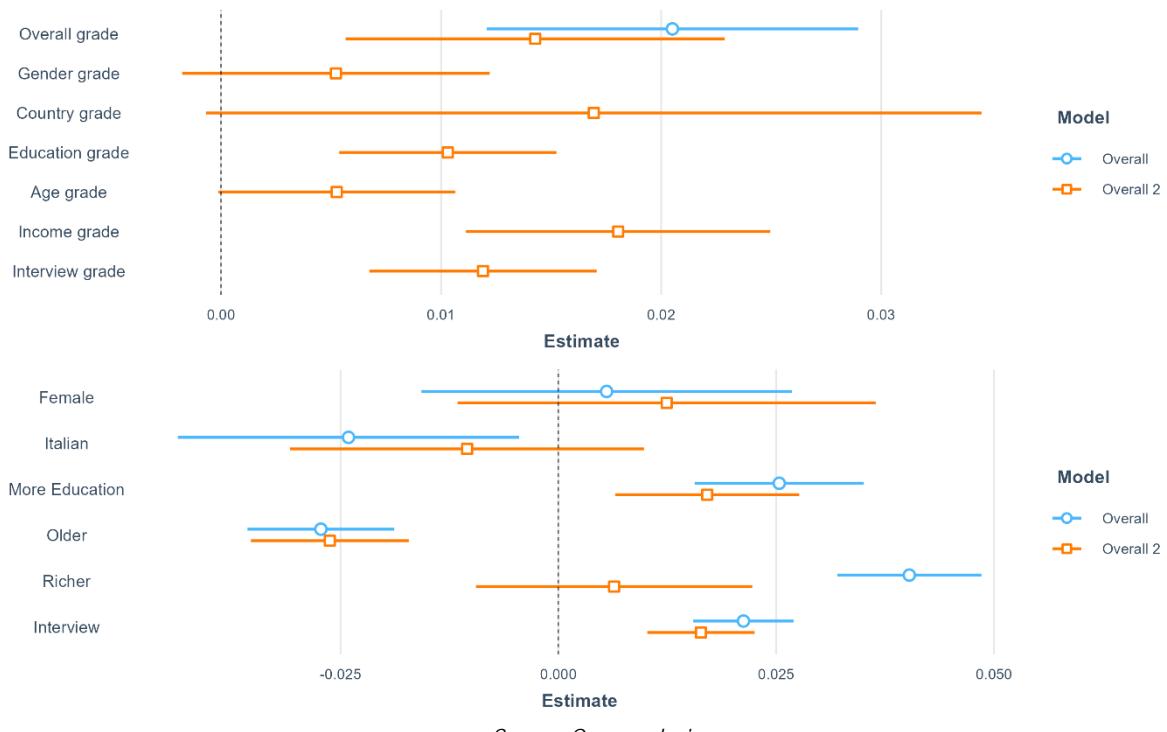
Source: Own analysis

We proceed to test how the AI system's recommendations, and their explanations in terms of grades for each applicant's characteristics, influence decisions. Figure 13 shows results of

regressions that confirm the positive impact of a better overall grade on selection of an applicant.<sup>17</sup> We also confirm that grades given by a fair AI are not more influential than those given by a discriminatory AI.

We further extend our analysis of the impact of grades given by the AI to consider grades on individual characteristics. We focus in particular on differences in grades by country and by gender, which only appear when the AI was discriminatory (the fair AI rated both dimensions as irrelevant and thus did not grade them). This also allows us to test whether decisions move in the direction of discrimination suggested by a discriminatory AI. We find that grades on gender are only marginally influential in biasing choice in favour of men (remember that the discriminatory AI gave better grades to men than to women). However, we also see that grades on nationality do bias choice in favour of Germans. The level of influence of grades on other dimensions appear to be the same for both fair and discriminatory AI.

Figure 13 Influence of AI grades on choice.



Source: Own analysis

All this analysis is done controlling for the characteristics of applicants, rather than just the grades given to them by the AI. Indeed, we want to exclude a possible explanation for the influence of the grades of the AI, which would be that the AI graded applicants in a way that is consistent with deciders' preferences. The effect of grades on selection are robust to those individual controls. This confirms that the AI has an independent effect on selection. We also note from a comparison with Figure 11 that the effect of applicants' characteristics on selection is very similar to their effect in the absence of an AI.

<sup>17</sup> The estimation equation is of the form  $\text{choice\_1} = \text{overall\_grade\_diff} + \text{gender\_grade\_diff} + \text{country\_grade\_diff} + \text{education\_grade\_diff} + \text{age\_grade\_diff} + \text{income\_grade\_diff} + \text{interview\_grade\_diff} + \text{W\_vs\_M} + \text{Ita\_vs\_Ger} + \text{education\_diff} + \text{age\_diff} + \text{income\_diff} + \text{interview\_diff}$  whereby  $\text{grade\_diff}$  is the difference in grade given by the AI between applicant 1 and applicant 2.

### 5.2.3 Effect of AI on discriminatory outcomes

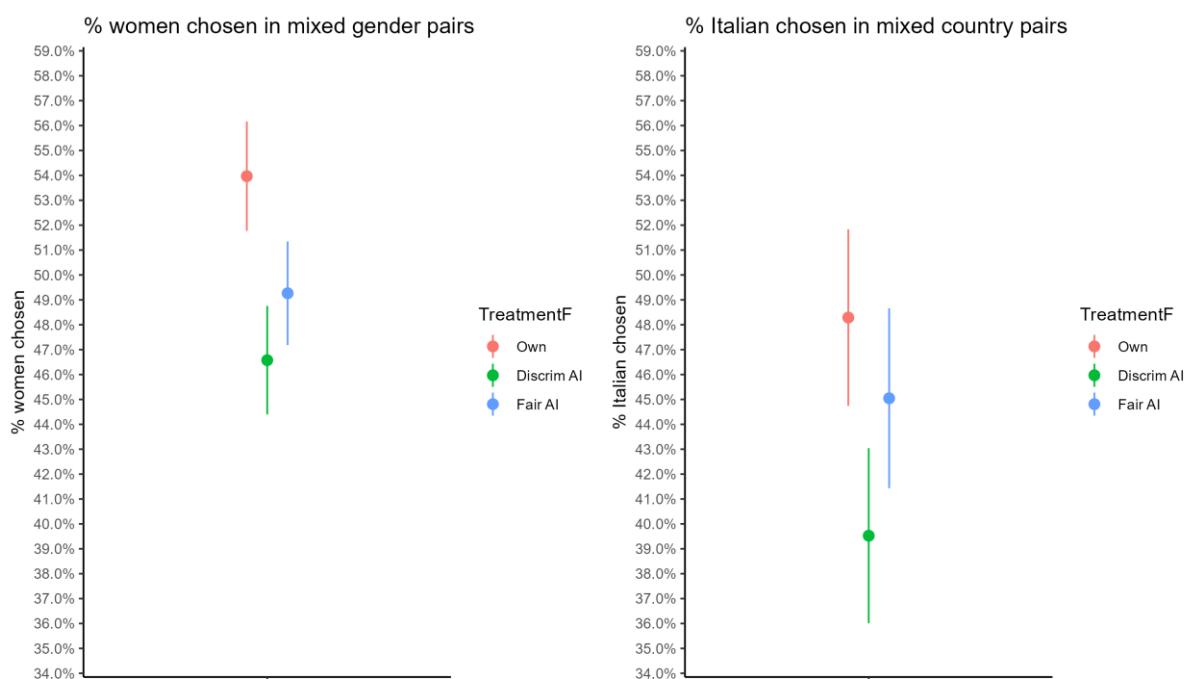
We saw in the last two parts that there is discrimination without AI and it is driven by deciders' preferences across candidates. We also saw that AI recommendations are influential, equally so whether the AI is fair or discriminatory. We further consider the overall impact of AIs on the level of discrimination across applicant. We run the same regressions as in column 1 of Table 24 in Annex 5, which dealt with choice without AI, to now consider choice with AI. We thus relate levels of discrimination by gender and country with the type of AI used. We summarize results in Figure 14.<sup>18</sup>

We find that gender discrimination is reduced when using either a discriminatory or a fair AI, whereby men are less discriminated against. Discrimination by country is increased when discriminatory AI is used, whereby Italian applicants are less likely to be chosen when the discriminatory AI is used.

The explanation for reduced gender discrimination is different depending on the AI used. The discriminatory AI recommends men by giving them a good grade, while the fair AI does not grade by gender. In both cases, this has the effect of countering an existing bias for women.

In terms of country discrimination, there is no significant overall tendency to favour Italians, so the fair AI which also does not differentiate by country does not change the situation. On the other hand, the discriminatory AI which grades Germans higher than Italians influences choice against Italians.

Figure 14 Gender and country discrimination, by AI type.

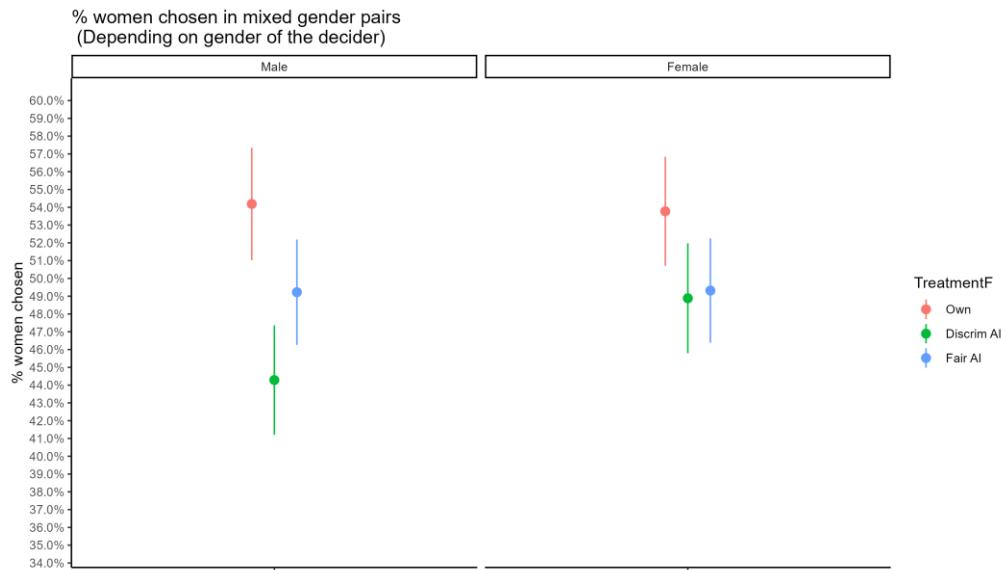


Source: Own analysis

<sup>18</sup> The estimation is based on a linear random effects regressions of the form  $\text{choice\_1} = (\text{W\_vs\_M} + \text{Ita\_vs\_Ger} + \text{education\_diff} + \text{age\_diff} + \text{income\_diff} + \text{interview\_diff})$ , as before, which we interact with the treatment variable indicating the type of AI (none, fair, generic). Figures show ex-post average estimates.

We saw in the previous parts that preference in terms of gender and country of the applicant depended on the gender and country of the decider. We therefore proceed with an investigation of the interaction between deciders and AI preferences by considering the effect of AI on gender and country discrimination depending on the gender and country of the decider, controlling for other characteristics of the applicants (Figure 15 and Figure 16).<sup>19</sup>

Figure 15 Effect of AIs on gender discrimination by gender of decider



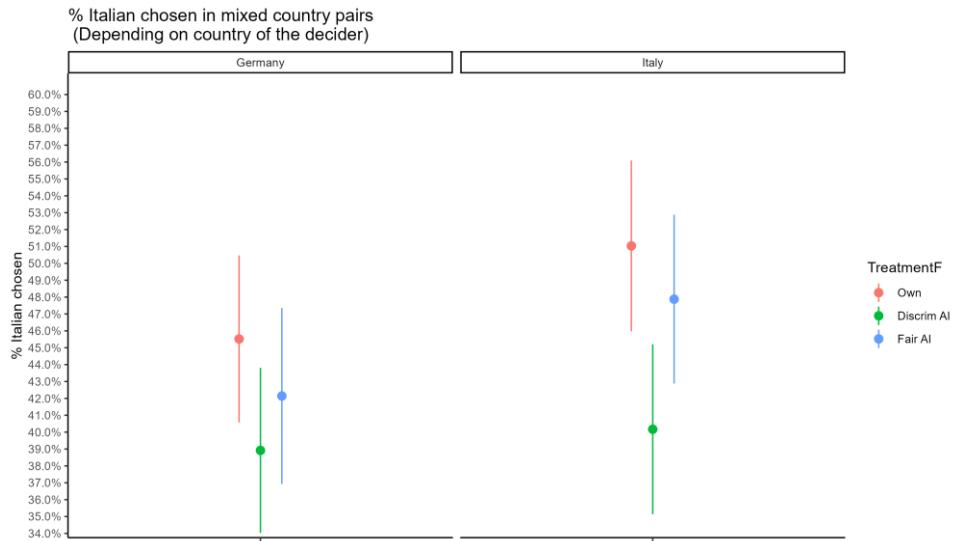
Source: Own analysis

We see in Figure 15 that the discriminatory AI, which favours men, changes a bias against men when deciders are on their own into a bias against women, among male deciders. The fair AI on the other hand reduces bias against women among men into no gender bias being apparent, as both genders are as likely to be selected. The effect of the discriminatory AI is not as pronounced for women, as it reduces the % of women selected, but only to a “fair” level (50%).

Figure 16 shows the same analysis depending on the nationality of the decider. A bias against Italians, which is apparent for deciders in both countries, is amplified by the same bias against Italians in the discriminatory AI. The magnitude of this change is more pronounced among Italians, who unlike Germans were not biased against Italians when without AI support. The fair AI does not change bias against Italians among German deciders, even though the fair AI itself is not biased by nationality.

<sup>19</sup> Estimation equations are of the same form as in footnote 18, except we now add a further level of interaction with the gender and the country of the decider.

Figure 16: Effect of AIs on country discrimination by country of the decider



Source: Own analysis

#### 5.2.4 Interaction between individual and AI biases

We saw that fair AI does appear to reduce gender discrimination, but that DMs's preferences also plays a role. Indeed, an AI that favoured men influenced men more than women to favour men, and an AI that favoured Germans influenced Germans more than Italians to favour Germans. We further our investigation of hypotheses 9 and 10 by relating expressed preference for a type of applicant with choice of this type of applicant, depending on the AI used.

We also investigate the possibility that support from an AI system allows deciders to discriminate in favour of their preferred type of applicant more precisely, as hypothesized in (T. A. Khan 2023).

We thus relate the likelihood to choose an applicant of a given type to the preferences among applicants expressed by the deciders before making decisions.<sup>20</sup> We report results of regressions interacting those two factors in Figure 17 for gender and nationality.<sup>21</sup>

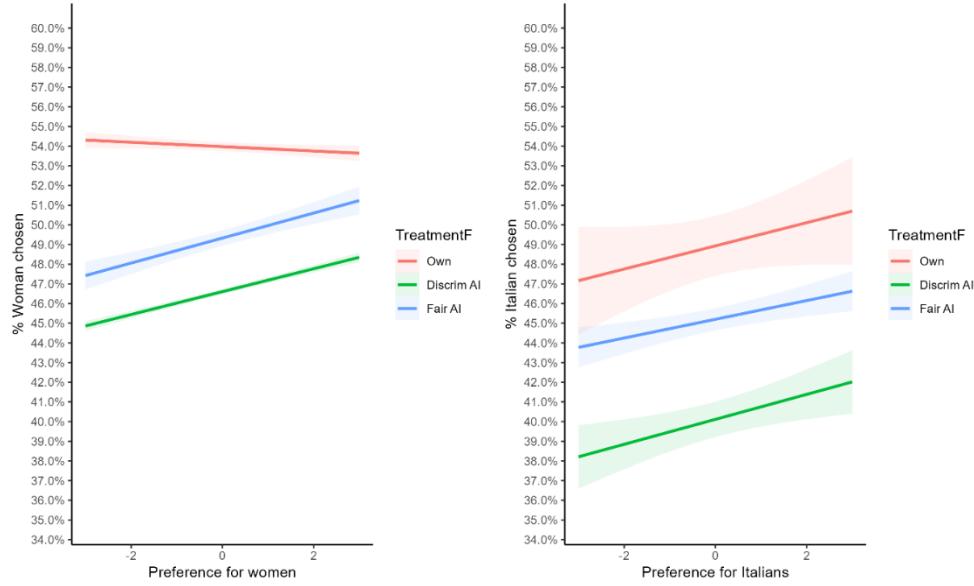
---

<sup>20</sup> Preferences are computed as explained in section 5.1.4 for each dimensions in the characteristics of applicants.

<sup>21</sup> The estimation is based on a linear random effects regressions of the form

$\text{choice\_1} = (\text{W\_vs\_M} * \text{Preference\_Women} + \text{Ita\_vs\_Ger} * \text{Preference\_Italy}) * \text{Treatment} + \text{age\_diff} + \text{education\_diff} + \text{income\_diff} + \text{interview\_diff}$ . Figures show ex-post average estimates.

Figure 17 Bias as a function of own discriminatory preferences, for gender and nationality.



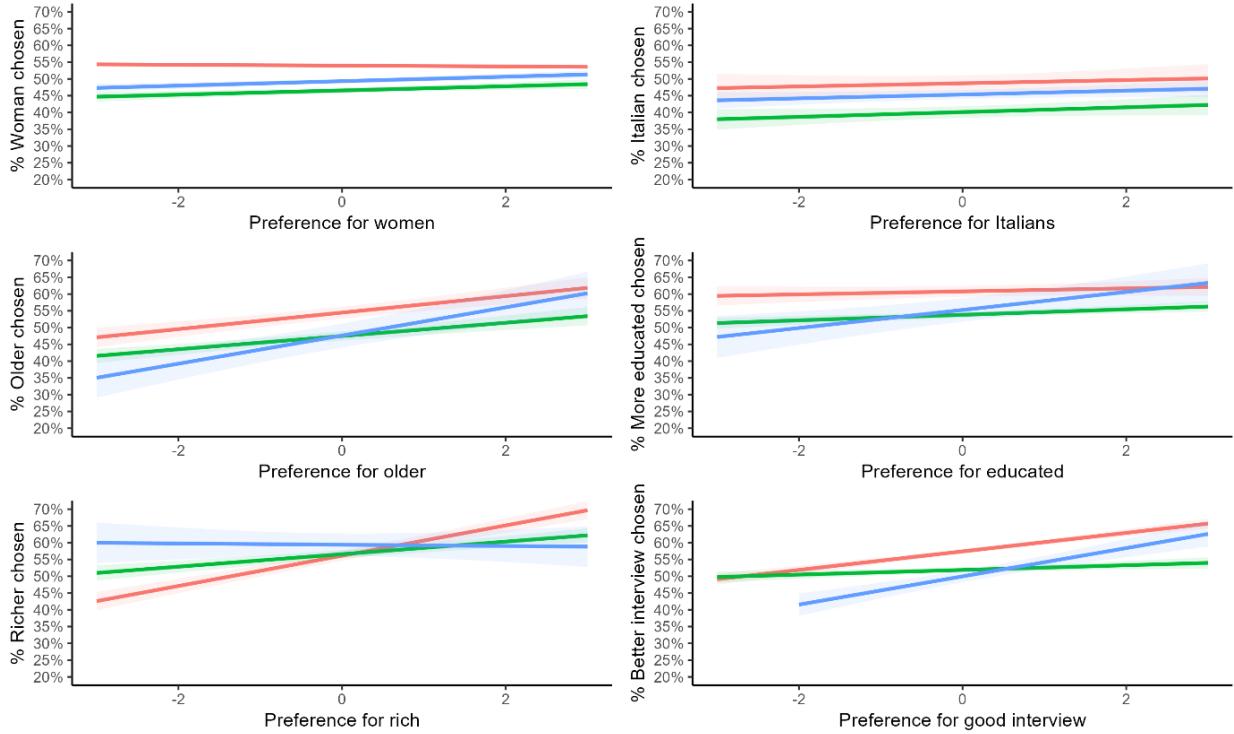
Source: Own analysis

The results shown in Figure 17 are consistent with those shown in Figure 14 in terms of average % of women and Italians chosen. The figure completes those results, along with their analysis by country and gender, to confirm that the decider's preferences also have an impact (% of women and Italians chosen depends on the expressed preferences of the deciders). The more a decider expressed preference for Italians/Women, the more likely he or she is to choose an Italian/Woman. Importantly, the magnitude of the influence of those preferences (the slope) is not consistently higher or lower depending on whether AI support was given or not, or depending on the AI. Indeed, lines in Figure 17 are parallel in most cases, meaning that individual preferences do not have more of an influence on choice when there is an AI or none.

We complete this investigation by considering the impact of other preferences on choice (Figure 18).<sup>22</sup> We confirm there is no clear pattern whereby AI DSS would make choice more sensitive to participants' characteristics.

<sup>22</sup> The estimation equation is of the form  
 $\text{choice\_1} = (\text{W\_vs\_M} * \text{Preference\_Women} + \text{Ita\_vs\_Ger} * \text{Preference\_Italy} + \text{age\_diff} * \text{Preference\_Old} + \text{education\_diff} * \text{Preference\_Educated} + \text{income\_diff} * \text{Preference\_Rich} + \text{interview\_diff} * \text{Preference\_Interview}) * \text{TreatmentF}$ .

Figure 18 Bias as a function of own discriminatory preferences, for all applicant characteristics.



Source: Own analysis

### 5.2.5 Differences in the perception of the AI

In this last section, we investigate deciders' perceptions of the different AIs, elicited after the experiment. We report results of regressions relating how far deciders reported they relied on and understood the AI, as well as whether they perceived it as fair or accurate (Table 17). We find no differences in perception depending on whether the AI was fair or not, whether in HR or in Banking. Women and deciders in Banking consistently report lower levels of agreement with the statements.

Table 17 Perceptions of the AI by deciders

	Dependent variable:			
	Relied	Understood	Fair	Accurate
	(1)	(2)	(3)	(4)
Female	-0.188**	-0.142*	-0.196***	-0.184**
Italy	-0.052	0.076	-0.012	0.051
Banking	-0.297***	-0.315***	-0.259**	-0.258*
Fair AI * HR	0.080	0.016	0.141	0.115
Fair AI * Banking	0.109	-0.017	0.030	0.028
Constant	2.915***	3.146***	3.068***	3.052***
Observations	951	951	951	951
R <sup>2</sup>	0.037	0.043	0.043	0.040
Adjusted R <sup>2</sup>	0.032	0.038	0.038	0.035
Residual Std. Error	0.910	0.870	0.904	0.897
F Statistic	7.277**	8.518***	8.514***	7.871***

Note: \*p<0.05; \*\*p<0.01; \*\*\*p<0.001

Source: Own analysis

Participants who were exposed to fair or to generic AI do not report differences in their perceptions of them, even in terms of fairness. This is even though we clearly stated the nature of their AI when we explained the AI and its recommendations to them. We further considered however whether participants' perceptions of the AI, and the importance for them to make fair decisions, follow their instinct, or make correct decisions, influenced their likelihood to choose the AI's recommended candidate.<sup>23</sup> We find no significant relations, even between reporting to have relied on the AI and the likelihood to follow its recommendations. This underlines the difficulty in relying on self-reports by deciders to evaluate their decisions, and their apparent lack of insights into their own decisions. This type of finding is typical in decision-making research and psychology. Some of this issue is due to the complexity of the decisions that participants had to make, especially in a setting that was unfamiliar to some of them. Other issues that come into play are a social desirability bias (reporting behaviour that will be viewed favourably by others) and the related experimenter demand effect (reporting behaviour along what one thinks the experimenter wants to hear).

We now move from an analysis of the quantitative stage of our mixed methods research to an analysis of its qualitative stage. We aim to go deeper into the complexities and subtleties of the context of our experiment, as experienced by our participants. This is something that is difficult to capture with numbers alone. Through employing semi-structured interviews, focus groups, and observational techniques, we seek to enrich our statistical findings with the nuanced perspectives and experiences of our participants. This qualitative exploration provides the depth needed to construct a better understanding of the phenomena under study.

<sup>23</sup> The estimation equation were of the form choice\_1=overall\_grade\_diff\*(AIRelied+AIFair+AIUnderstood+AIAccurate) and choice\_1=overall\_grade\_diff\*(Importance of fairness+Reliance on instinct+Goal to be correct).

## 6 Results of the study's qualitative part (Qualitative)

### 6.1 Sample descriptive statistics.

#### 6.1.1 Participant for the Individual Interviews

We invited a subset of the sample to participate to individual interviews. After the recruiting phase, we concluded with (N=13) participants with whom we organized interviews. The selection of participants was based on their profession (human resources, finance) and their nationality (German, Italian). Table 18 presents the participants demographics. Four out of thirteen participants were German all of which worked in Human Resources and nine Italian, five of which worked in Human Resources. Despite our targeted recruitment strategy, German participants were less responsive which resulted in a larger number of Italian participants. However, the intended diversity of the sample was reached considering other demographics, such as age, gender, and size of the company each participant was working at the time of the interview's conduction.

Table 18 Distribution of interview participants

	Human Resources (HR)	Banking (FIN)
German (DE)	4	0
Italian (IT)	5	4

Source: Own analysis

In the following, we will code citations from participants using codes shown in Table 19. The code states the country (IT=Italy, DE=Germany) and profession (HR=Human Resources, FIN=Banking) of the participant, and a number from 1 to 13. The table gives further details on the participant (gender and age range).

Table 19 Demographics of interview participants, with participants codes

CODE	GENDER	Age-range
IT-HR-01	Male	46-55
DE-FIN-02	Female	56-65
DE-FIN-03	Male	31-45
DE-FIN-04	Female	31-45
IT-FIN-05	Male	31-45
IT-FIN-06	Female	31-45
IT-FIN-07	Female	31-45
IT-FIN-08	Male	31-45
IT-HR-09	Male	31-45
IT-HR-10	Female	46-55
DE-FIN-11	Female	31-45
DE-FIN-12	Male	46-55
IT-HR-13	Male	31-45

Source: Own analysis

The interviews took place online in participants' native language (German or Italian) through video-calls. Participants received a compensation of 20 euros per hour for the time spent for the interviews and the workshops. Each interview lasted 1 hour, and each workshop 2 hours.

### 6.1.2 Participants for the group workshops with professionals

We invited all the individuals that participated to the individual interviews to join the workshops. After the recruiting and interviewing phase, not all the participants were interested or available to participate to the follow-up workshop. Despite our targeted recruitment strategy, German participants were, once more, less responsive which resulted in no workshops being delivered in German. However, the results presented in the Italian workshops were also encompassing the overall and German results, and we managed to obtain the intended diversity of the sample for the workshop that took place, considering other demographics, such as age, gender, and size of the company each participant was working at the time of the workshops' conduction.

We thus concluded with 2 workshops, (one for HR and one for FIN), with a total of (N=7) participants. Table 20 presents the participants' demographics. All the workshop's participants had been previously interviewed.

Table 20 Distribution of workshop participants

	Human Resources (HR)	Finance (FIN)
German (DE)	0	0
Italian (IT)	3	4

*Source: Own analysis*

Due to the conversation being in groups and interactive, also using the support of visuals, it is not possible to trace back the quote to the individual participant. Some of the considerations, opinions and quotes were the result of a collaborative reflection, where multiple participants build on the other initial points. We therefore label the quotes only using the code for identifying the nationality and the sector.

### 6.1.3 Participants in the workshop with experts

Regarding the workshop with the expert, we tried to have a diverse team of experts. We selected the experts from the fields of AI, ethics, design, art, science, philosophy, and sociology to guarantee a diverse representation of perspectives, ideas, and visions on the topic. Table 21 lists our participants, and their specialization / discipline.

Table 21 List of Experts with their specialization/discipline.

Name	Discipline
Valeria Adani	Designer, Specialised in Trust and AI
Christina Melander	Designer, specialised in Tech and Ethics
Suhair Khan	Technologist and Designer
Raziye Buse Çetin	AI and ethics researcher
Filippo Cuttica	Designer, artist and researcher, specialised in Ethics
Abdelrahman Hassan	Creative technologist, AI Strategist , and Digital ethicist
Tim de Jonge	Researcher in Artificial Intelligence and Fairness
Emanuel Dietrich	Fairness in Distributed Systems
Senka Krivic	Researcher in explainability in AI and machine learning
Egon van Born	Data Scientist
Giada Pastilli	Ethicist and Philosopher

We also invited inhouse specialists in behavioural insight, design, and foresight from the EU Policy Lab to contribute to the organisation and discussion.

## 6.2 Analysis of results

This qualitative study generated three data sets based on (i) the individual interviews with professionals, (ii) the online group workshops with professionals, and (iii) the workshop with experts. Below we describe the three data sets:

- Interviews: Individual verbal conversations with interviewer (member of the research team) which were videorecorded and automatically transcribed and translated with the use of Microsoft OneDrive. The data then, were manually annotated according to the questions of the interview and were inserted into an Excel file. The data then received a second iteration of annotation, directly on the Excel file, according to the research questions RQ1, RQ2, and RQ3.
- Workshops with the participants: Written material on the MIRO application in combination with the researcher's live and shared notes based on the online conversations with the groups of participants. The material was discussed with the research team and was organized and annotated on the MIRO application.
- Workshop with the experts: Participants' notes, and paper-based materials from the physical workshop. The material included sticky notes with short texts, longer texts, pictorials, and other visuals. The physical material was digitalized after the workshop by members of the research team who were present in the workshop and were analysed based on the research question RQ4.

This data was coded, which involves assigning labels to segments of the data to summarize or categorize them. The process of coding is a core part of the qualitative data analysis (Flick 2013). This process helps to identify patterns and themes in the data, laying the groundwork for subsequent data interpretation and presentation. Qualitative research often involves multiple iterations of coding, creating new and meaningful codes while discarding the ones that do not serve the research questions, to generate structures meaningful for the research goals.

As such, the following sections present the way that each of the three categories of data were cleaned and organized, annotated, and reorganized to be interpreted through the lenses of the research team. We acknowledge that, as per the qualitative research paradigm, subjectivity of the qualitative results as being interpreted by the research team is inherent in this study; however,

since this research team has been on purpose created with a highly multidisciplinary approach, internally we iterated an interrogated the results of the study with a critical view from the perspective of multiple disciplines.

### 6.2.1 Results from interviews with professionals

We followed an iterative procedure for our data analysis according to the framework proposed by (Srivastava and Hopwood 2009) that suggests iterations of the data analysis not as a mechanical procedure but as a tool for reflection and further understanding of the data. As such, we first annotated the data according to the interview questions; then according to the higher-level research questions of this qualitative study; finally, we integrated the two analyses (Table 22).

Table 22 Iterative procedure of data analysis in three steps.

- 
1. Analysis of the data based on the initial questions of the interview
  2. Analysis of the interview data based on higher-level research questions
  3. Integration of the two analyses – reflection on AI and Human Oversight
- 

We employed a combined inductive and deductive qualitative approach for data analysis (Strauss 1987). We initially created a codebook through inductive open coding, considering the RQs and the existing literature, with topics related to our theoretical framework. Our starting point was the EU ethical guidelines on AI. We automatically transcribed and translated all the videos with the interviews and the material was prepared for manual annotation.

Before the data analysis, two members of the research team discussed the thirteen transcripts of the interviews. One member of the team manually annotated the data based on the questions of the interview. Two interviews were independently annotated by a second member of the team and examined for inter-rated agreement. After a comparison of the indecently annotated interviews, an inter-rated agreement of 87,5% was found. The inconsistencies were discussed in a team meeting and differences were resolved. The responses for those questions of all transcripts were re-examined to ensure consistency in the annotations.

While the annotations in the first iteration which involved participants' answers to specific questions were relatively structured and straightforward, the annotation in the second iteration required critical reflections from the research team to identify the emergence of specific topics. As mentioned, the guiding starting point was the research questions of the study. However, often, participants would create informative narratives that the research team decided to include, even if this was an outlier. As such, while the main strategy for the inclusion of a theme was its frequency among the participants, the relevance of a theme would also function as a criterion for the inclusion of the specific topic in the results of the study. For this reason, the research team decided not to present descriptive statistics of the results; rather we present the results as a critical reflection that could potentially inform the results of the quantitative study and trigger more questions about the follow-up study with the policymakers.

### 6.2.2 Iteration 1: Analysis based on the initial questions of the interviews

As described in section 4.2.1, the interview questions focused on three topics: (1) contextual information and current practices on the use of AI, (2) interaction with a biased/unbiased system, and (3) reflection on priorities and biases in decision making in the specific scenarios.

Below, we present the results of the interviews based on the corresponding questions for each topic.

## A. Contextual information and current practices on the use of AI

In this section, we present additional contextual information about the participants and their current practices on AI.

a. **Participants' demographics:** Because of the relatively small size of the sample, we examined the demographics diversity. Previous research shows that demographics correlate with the use of AI in the workplace. For example, a survey by EUROSTAT showed that large enterprises are more likely to use AI than medium or small ones.<sup>24</sup> Below we provide an overview of the demographics we collected during the interviews:

- Age-group: As shown in Table 18, the age group of most of the participants was 31-45, with three participants in the age group of 46-55 and one participant falling into the age group of 55-65. This allocation indicates that the participants for the qualitative study were diverse in terms of age groups which would allow us to explore the responses and participants' attitudes against their age.
- Geographical: While we controlled for participants' nationality (Italy and Germany), we observed that participants' place of origin was relatively homogeneous in terms of the size of the place they lived with most of the participants living in mid-sized places, such as Dusseldorf and Milan. Three participants lived in small towns.
- Size of organization: The size of the organization in which the participants of the study were employed varied a lot with one of the participants working for the family company which employed 3 people while another participant worked for an organization that employed more than 600K employees. However, most of our sample worked for middle-sized companies of 50 to 100 employees.
- Position and role within the organization: All participants were relevant to positions that dealt with customers in various ways and their role included processes of decision-making either in a direct way (interacting with candidates in HR or loans) or in an indirect way by supporting decision-making processes often including automated systems.
- Time of experience in the specific company: Participants differ largely in the duration of their employment at the specific companies.

b. Current practices on the use of AI: With questions 7 to 13 of Table 5, we allowed participants to talk about their current practices in using AI systems for their daily jobs. In their descriptions, we observed that often participants talked about automated systems or applications that currently facilitated their tasks which are not necessarily based on AI. However, all participants were able to make the distinction between applications that are used for the organization of tasks such as the organization of candidates' demographics or other information related to them and systems that function as recommenders or decision-making support. Below we highlight some examples that are relevant to the main goal of this study which is the exploration of human oversight in AI-supported decision-making.

- Classification of AI systems: Participants indicated several uses of AI in their everyday work activities, such as an automatic organization tool for databases (DE-FIN-11, IT-FIN-07), administration tool (IT-HR-12), a tool for financial predictions ((DE-FIN-03), as a tool for data processing and analysis, as a tool for communication with the client (DE-FIN-11), data visualization (IT-HR-01), skimming of data (IT-HR-09), writing texts with generative AI applications (DE-FIN-04). However, in some cases, we observed some confusion about classifying a tool as an AI or not. For example, participant DE-FIN-11 mentioned that while they use tools for automatic decision-making, she would not call these tools AI. More specifically, she mentions: *"And now in my own team, we just use Python [...] More or less a decision tree. I don't know if you really have to see that under AI."* [DE-FIN-11].

---

<sup>24</sup> [https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Use\\_of\\_artificial\\_intelligence\\_in\\_enterprises#Enterprises\\_using\\_artificial\\_intelligence\\_technologies](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Use_of_artificial_intelligence_in_enterprises#Enterprises_using_artificial_intelligence_technologies)

- The usefulness of AI systems: Participants indicated that some activities in their daily work would not be possible without the use of AI. For example, participant IT-HR-01 mentioned: *"Broken down by categories and manual processing of these data would not be impossible, it would be very complicated"* [IT-HR-01]. However, some other participants mentioned that AI tools are not ready yet to perform certain tasks that relate to communication with the customer: *"Personally, however, I think that the AI is not yet so exorbitantly mature that to make a decision about a candidate, there is really the human component missing"* [DE-FIN-03].
  - Importance and complexity of the task: Across participants, there was a tendency to reflect on the use of AI tools in relation to the importance and complexity of the specific task at hand. Participants for complex and important tasks would be more reluctant to rely on AI tools. For example, participant DE-FIN-03 mentioned: *"About Artificial Intelligence: Will certainly help us, but I hope that especially in the field of financial services, the topic is so complex that it will be a bit more difficult with artificial intelligence to take away our jobs"*. Other participants mentioned that they would use or not use AI depending on what kind of risks this would have. For example, participant IT-FIN-06 mentioned *"In the finance sector, I would not let AI to make a decision for me; putting the money in the hands of an artificial intelligence? Better not"*. [IT-FIN-06]
  - Collaboration with the system and justification of AI recommendation: Some participants mentioned that they partially use AI for the selection of candidates. They indicated that they consider the AI recommendations but, in the end, they would check about its credibility. For example, participant DE-FIN-04 said: *"Say that we have a 40-year-old who asks for a loan for his company; how should I assess the risk? Some AIs calculate and suggest. We can stick to them or it's just a good recommendation. We are told so, but if we have a good justification and say that I have spoken to the customer and I assess their credibility in a specific way, then we may deviate from the AI recommendation"* [DE-FIN-04]. This is a typical example of how an employee would use AI as a support system and it indicates the employee's need to be able to make the final decision. This applies in the cases when our participants talk about the decision-making process without considering other contextual factors, as will be discussed later in this chapter.
- c. AI, ethics, and policies of their institutions: With question 14 of Table 5, we were interested in understanding if the company in which the participants were employed had some centralized ethical procedure or policy regarding the recruitment of candidates in both our scenarios to prevent discrimination. The responses we received were mixed with a general emphasis on the issues of data privacy. Most participants were not aware of whether their company had an organized way to prevent discrimination, especially by AI systems. Others would indicate that the instructions they receive from the hierarchy of their company focus mainly on the performance of the candidate. For example, participant IT-HR-12 says, *"For internal competitions, I imagine that what matters in internal calls for various types of positions is not the gender, experience, and resumé are important"* [IT-HR-12]. Similarly, participant IT-HRI-07 said: *"Skills are the most important criterion, rather than looking if the candidate is a male or female. We always seek, for example, competence, experience, and quality"* [IT-HRI-07].

## B. Interaction with biased/unbiased system

The second phase of the interview focused on instances of the experimental study and participants' rationale when interacting with a biased or unbiased AI system. With questions 1 and 2, we asked participants to choose the most suitable candidate between two with different demographics while thinking aloud and explaining their choice. This led to the following insights.

- a. **Inter-**relations of candidates' characteristics:**** For the selection of a candidate, most participants would initially consider all the demographical information of the candidate and they would proceed to identify relationships among those demographics and additional characteristics, such as interview performance. For example, participant IT-FIN-06 reflected as follows: *"Man or*

*woman does not give me a lot of information. Age similarly. However, this German is more towards retirement age, compared to the Italian woman. In terms of nationality, I can assume slightly better working conditions in Germany than in Italy in terms of wages, etc. Maybe the German also has a job position that allows you to earn better. Educational level: high, medium. I think he's a man who has a work experience, he's reached a certain level. Let's say the assessment I made gives importance to age but not to nationality*" [IT-FIN-06]. As such, in this quotation, it is evident that the participant does not only consider the characteristics of the candidate in a linear and summative way, but they try to find inter-relations and interactions among the given characteristics (see footnotes 25 and 26 on page 60 for some considerations on this).

- b. Contextualization and situatedness of candidates: During the phase of reflection, we observed that often participants examined the provided characteristics for the candidate, in relation to the wider context of the participant and the current situation. One of the elements that emerged among the participants was the examination of the characteristics of the candidate in line with the strategy of their company. "*I have to contextualize this screen. With the criteria of my company*" [IT-HR-12]. This element of prioritization of the strategy of the company appears in some participants in the first phase of the interview, which might relate to the trust of their company or the authority of the company to "influence" certain decisions". Interestingly, the participants would not question these strategic indications in the decision-making process.
- c. Explainability and reinforcement: After the first interactions with the (biased and unbiased) AI, the interviewer revealed the reasoning behind the algorithmic recommendation and invited the participants to reflect on it. Some participants expressed their need to see the reasoning of the algorithm in earlier stages of the interview, too. "*The attribution of a score dictated to me by an artificial intelligence in a way that was not entirely clear*" [IT-HR-01]. Another participant mentioned explicitly "*I would like to see the algorithm*" [IT-HR-12]. Some participants talked about the fact that behind the algorithmic recommendations, there are human decisions that are not revealed when they should be. For example, participant [DE-FIN-04] mentioned: "*Nevertheless, it's a decision support, but you should still always see the real person behind it. And yes, don't make this decision based solely on these facts*" [DE-FIN-04]. Interestingly, one of the reasons that some participants would appreciate the introduction of explanations of the recommendations of the AI system is the possibility of the reassurance and reinforcement of their (human) result of a decision.
- d. Transparency: As a last part of this phase, participants would be informed about the way the specific algorithms were trained. As such, in addition to the introduction of explanations, some participants required to know more about the foundation of the specific recommendations seeking to trust them more as well as the possibility that the algorithm learns by its interaction with the human decision-maker. "*But it's figured out my logic by now, so it already knows what I'm going to choose.*" [IT-HR-12].

#### C. Reflection on priorities/ biases in the specific scenarios

The last part of the interview intended to understand the participants' reasoning and possible criticism about the decisions provided by the AI system.

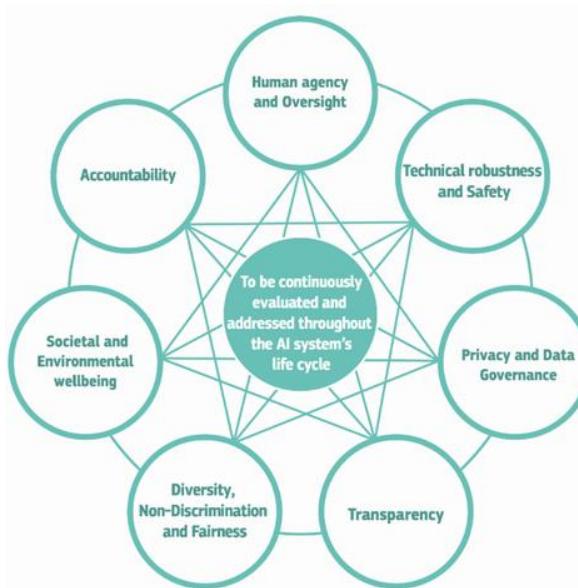
- a. Contextualization: Most participants replied by considering the context of their decision making. "*From a rational point of view and from a statistical point of view, the decision is absolutely understandable. Nationality and gender from the human side, but then again from the economic side it would be important. From an economic point of view, I think nationality should not be ignored. Gender can go away completely. Because right now, Italy's creditworthiness is not exactly that high and when I talk to customers about short-term investments or when they say, they just want to park money in the money account then of course I say to the customers okay Sweden, Germany, Triple A.*" [DE-FIN-03].
- b. Ethical vs Pragmatic: It is worth noting that some participants would reflect on the AI decision not only in terms of non-discrimination, meaning from an ethical point of view but also from a pragmatic point of view. For example, participant IT-FIN-06 mentioned that the discriminatory recommendation by the AI was "*Less ethical, but reflecting reality*" [IT-FIN-06].

c. AI, under whose service? In addition to previous instances during the interview where participants would accept a discriminatory recommendation by the AI if this was embedded in the strategy of their company, in the last part of the interview some participants made the case that to assess a discriminatory decision as ethical or not, one should think of “*Whose perspective are we taking? The candidate or the bank? A bank is there to make money, not to help people.*” [DE-FIN-03].

### 6.2.3 Iteration 2: Analysis of the interview data based on higher-level research questions.

As mentioned in section 0 (methodology), after the first iteration of the data annotation and analysis, the research team performed a second iteration taking a higher-level perspective, while trying to address the initial research questions. In the second iteration, we performed a thematic content analysis (Vaismoradi et al. 2016). As such, we developed an ad-hoc annotation scheme. We took as a starting point the Assessment List for Trustworthy Artificial Intelligence (ALTAI) which was developed by the High-level Expert Group for Artificial Intelligence of the European Commission (Figure 19). ALTAI is a practical tool that helps business and organisations to self-assess the trustworthiness of their AI systems under development (High-Level Expert Group on Artificial Intelligence, 2020, <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html>).

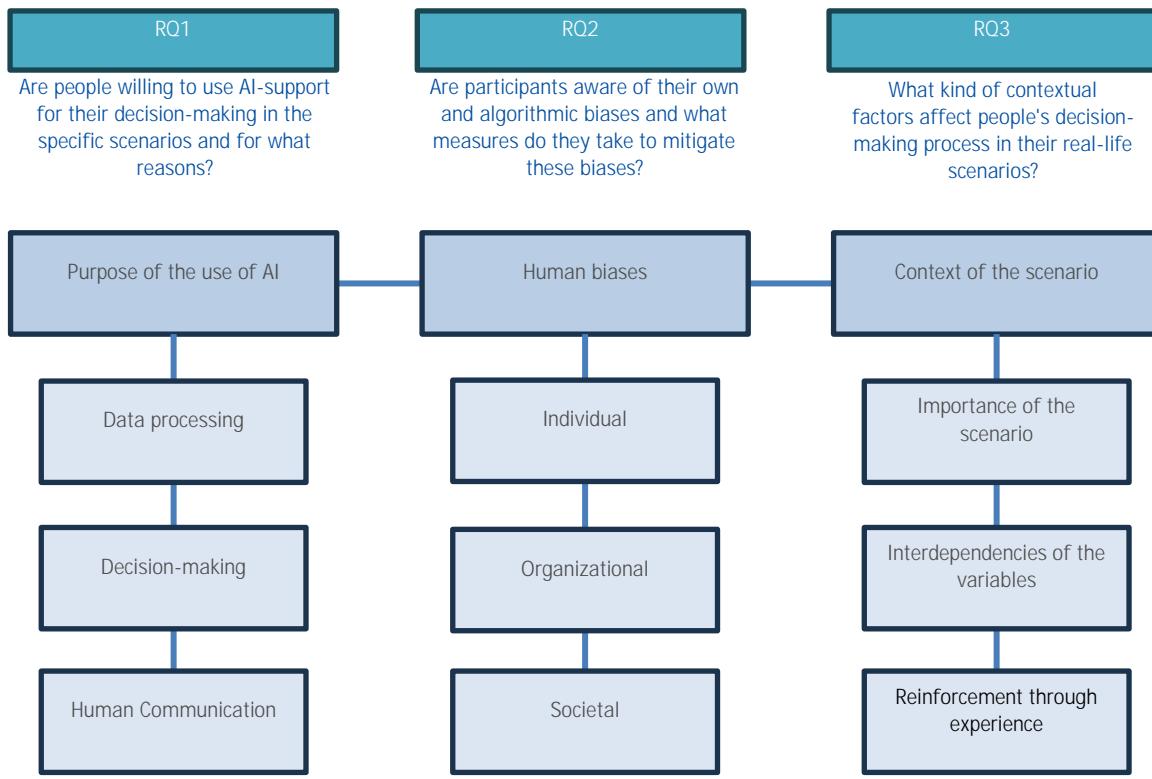
Figure 19 ALTAI Framework



Source: (High-Level Expert Group on Artificial Intelligence, 2020, <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html>)

Based on the above-mentioned ALTAI framework, we identified the topics that were relevant to our RQs. As such, the annotation scheme covered the general topics of (i) Human agency and oversight, (ii) Transparency, (iii) Diversity, non-discrimination, and Fairness. These were examined in relation to the RQs and additional subthemes were identified based on the responder's content. Figure 20 shows the annotation scheme in relation to the corresponding Research Questions.

Figure 20 Annotation scheme for the analysis of the interviews with the professionals.



Below, we present the main findings of the interviews analysis per Research Question according to this annotation scheme.

#### A. Research Question 1: Are people willing to use AI-support for their decision-making in the specific scenarios and for what reasons?

Throughout the interviews, participants indicated a positive attitude toward the use of AI for their professional activities. Participants talked about the use of AI tools as a tool that supports their jobs various ways, such as AI as a tool for data processing, as a tool for decision-making support and as a tool for human communication. Participants associated artificial intelligence with general words as "innovation, information, technology, automation", positive connotations as "reliability, potential", as well as connotation of "support, assistance" and "quality, learnings". Below, we list the three main categories of participants' views regarding the use of AI systems in their current jobs.

- a. AI as a tool for data processing Participants provided examples on their current practices of AI as a tool for data processing, analysis, classification of new data and data visualization. As such, algorithms are tools that help with quick, objective, and robust analysis of large amount of data; algorithmic support in this case is welcome by the participants and they see the added value without questioning the use of AI. *"An administrative software that takes care of the administrative accounting part that are not connected"* [IT-HR-12]; *"So, we're able to process thousands of analyses per hour, so to speak"* [IT-HR-12].
- b. Algorithms for decision-making support systems: In this case, algorithms do not only provide the results of data analysis, but also, they are used for decision-making with the recommendation of certain solutions. Participants expressed their concerns about the use of this kind of use of AI tool. Here, different perspectives might relate to the specific scenarios (HR vs Financial), for example, extraction of implicit information about the candidate. In the "Hiring" scenario, the importance of the interview was discussed extensively while in the "Loaning" scenario, participants prioritized "objective data" and, if their decision was on the borderline, then they would consider "interviews" and the algorithm's suggestion. *"Nuances, context, and data that*

*are not “writable”. We understand them only from the human interaction. [...] I disagree, not only because I don't know the criteria by which the algorithm makes these judgments, but also because of the merits of the judgments themselves* [IT-HR-01]. Similarly, *“In the end, there are certain criteria that are provided, the computer goes through the tree. I don't really want this as artificial intelligence. Of course, humans could do that just as well. Whereas with an AI, you are more likely to say that it comes to a decision that a human would sometimes not see, or perhaps would make differently. However, with this automated credit decision, the criteria are of course precisely specified and must be so that it can be understood. So, in this respect, the question then is, how do you really interpret AI?”* [DE-FIN-11]. With this participant's narrative, we can see that AI is initially questioned (*“I don't really want this as artificial intelligence”*), however, later, the very person mentions the limitations in human decision-making (*“with an AI, you are more likely to say that it come to a decision that a human would sometimes not see”*).

- c. AI as a tool for communication: Across participants, the factor of human communication with the candidate appears important and despite how objective is the algorithm, participants indicated that human communication is an element of value that cannot be replaced by the algorithm. Some of the reasons mentioned by participants are (i) Understanding a candidate's stimulus to work for a specific company and (ii) Obtaining more implicit information about the status of the candidate. Overall acceptance varies among participants; it also varies for the same participant over the time of the interview. For example, participant [DE-FIN-03] mentioned at the beginning of the interview: *“I find it a bit frightening that artificial intelligence is influenced by the gender of the candidate like that”* [DE-FIN-03]. While in the end of the interview, the same participant seemed to take the influence of human decision-making by algorithms as a mere fact *“Unfortunately, just hard, hard, hard facts”* [DE-FIN-03].

As such, the use of AI for the specific profession was categorized by the participants according to its function, which in turn, influenced their perceptions and acceptance. Participants talked about the use of AI tools as a tool that supports their jobs in at least three different ways: AI as a tool for data processing, as a tool for decision-making support and as a tool for human communication. Overall, the interview data show that the acceptance of an AI system in specific professions depends on the role and the capabilities of the system as well as on how pervasive the system is.

#### B. Research Question 2: Are participants aware of their own and algorithmic biases and what kind of measures do they take to mitigate these biases?

- a. Individual biases Participants take into consideration their own personal experience to make certain statements about the suitability of a candidate based on certain characteristic, such as the correlation of digital literacy with age. *“Younger people have some skills that we, at least I, have not be trained on, such as, familiarity and the use of computer science. My younger collaborators teach me things, but it is difficult for me, although I see that it is done faster. They teach me something which is natural for them but not for me. I learned in a different way and so I do it longer.”* [IT-HR-12].
- b. Organizational: Participants interrogated about the processes that might influence the algorithms and often allocated possible biases on their organizations. As such, they often considered that if their company takes certain decisions in favour or against certain characteristics of a candidate, the individual employee should adhere to these decisions. *“It depends on what our company policy says about it”*. [DE-FIN-04]. As far as Human Resources professionals are concerned, participant [IT-HR-01] said *“we must give priority to certain criteria in the recruitment for non-discrimination; this is now, let's say, rather an obligation in recruitment almost everywhere”* [IT-HR-01]. Some participants would prioritize the formal instructions they receive from their organizations even if this conflicts with the AI suggestion. The ultimate driver/decider is not always the person who is doing the selection activity. Biases and power dynamics rely also in the dynamic of the organisations the people are working in. When participants were asked in which occasions they would put their own doubts/judgement on the

- side, participants prioritized instructions from their institution, no matter the possibility of biases: “*If the organisation decides to hire a man for any reasons and if there are ‘valid motivations’, we will follow internal organisational policies or laws*” and that “*what is needed is the greater good even if it is not what I personally would like*” [DE-FIN-03]. These quotations indicate the role of the organizations in which the participants take decisions on how they take these decisions.
- c. Societal Participants often express their worries about general biases in society but when they were called to take a decision, they aligned their behaviour with the existing biases. “*Age talks about the flexibility and/or willingness to work and learn*” [DE-FIN-11].

### C. Research Question 3: What kind of contextual factors affect people's decision-making process in their real-life scenarios?

In the first iteration of our analysis, participants often talked about the contextual factors that affect people's decision-making process. In addition, they referred to instances that contextual factors would influence their tolerance and acceptance of a discriminatory recommendation by the AI. With this research question, we aimed to have an in-depth understanding about the role of the contextual factors on human oversight.

- a. Importance of the scenario: Participants highlighted that the importance of the specific scenario plays a substantial role in their interaction with the algorithm and their trust to the AI recommendation “*Putting the money in the hands of an artificial intelligence? Better not.*” [IT-FIN-06]. Participants answers indicate allocation of different degrees of urgency in terms of the contexts within which they should make a decision and how this context would influence the acceptance or not of an AI recommendation. For example: “*I mean, we're not talking about how to paint the door of the company, we're talking about a hiring process, then?*” [IT-HR-01]. Similarly, “*Yes, it depends. We would probably still finance a car, but not a single-family home for 30 years*” [DE-FIN-02].
- b. Complexity and Interdependencies of the variables: During the interview, often participants referred to various characteristics of the candidates, and how the evaluation of these characteristics affects their decision as well as the acceptance or not of an AI recommendation which is based on the evaluation of specific characteristics. However, participants indicated that the isolated assessment of the characteristics might be a simplistic way for the selection of a candidate and that special attention should be given to the connections and the interdependencies of the variables.<sup>25</sup> <sup>26</sup> As such, participants indicated that when humans take into consideration certain data for the evaluation of a candidate, these data have an aggregated form and different parameters which influence each other are taken into consideration. Many participants expressed their doubts about the ability of AI to go through such a process and if that was the case, they expressed their demand to know about it in order to make more informed decisions: “*And that's why it's always important to me to understand and understand exactly what AI does. It may very well be that the AI comes to different decisions and of course you can look at that. That can also be right, no, so I, I'm definitely open to that, would also want to use AI there, but still humans should still do that again. the AI can't evaluate what it is doing, in order to somehow classify this decision or possibly to say to the AI, yes no, male, female can actually not be a criterion*” [DE-FIN-11].
- c. Missing information: While participants had a certain number of candidates' characteristics with their rating available, they often talked about the need for further contextual information. “*A single-family housing estate or rather social housing? Are you married or are you not married?*

---

<sup>25</sup> This finding, whereby participants incorrectly assume that the AI is only taking into account variables in a linear way, was encouraged by our use of the LIME explanatory framework (M. T. C. Ribeiro [2016] 2024) to give explanations to participants about recommendations made by the algorithm. This simplifies complex non-linear global models into simple local linear explanations. This then gives the impression that those explanations represent what the model does.

<sup>26</sup> We note the relation between this finding and the finding in (Orfanoudaki et al. 2022) whereby humans tend to apply linear models of explanations. While participants say they are better able to take into account complex relationships than the AI, the opposite might be true.

*Such and such characteristics such as not married in housing, in social housing and so on and immediately gives deductions, so that you have a worse score than someone else, probably because it is assumed that someone who lives in social housing does not have as much money." [DE-FIN-02] Also, "I need additional information or justification in any way. I would need a resume to look for more contextual information about the profiles themselves.". [IT-HR-12]*<sup>27</sup>

## 6.2.4 Results from the workshops with professionals

Following the interviews, we organized small-group workshops with the participants to trigger further discussion and reflections. Workshops were held online in the native language of participants and were facilitated by the research team. The workshops were recorded for further transcription and analysis, together with the visual material produced during the workshops by the participants. We present here a thematic content analysis of the discussions:

### 6.2.4.1 Reinforcement through experience

Participants highlighted that experiencing for themselves the effectiveness of the AI recommendation would be one of the factors that would influence or have influenced their trust in the recommendation. A participant thus talked about their mental strategy to decide if a candidate was suited for a loan "*The more I use a certain strategy that is proven effective the more the good results I have, and then the more I use this strategy.*" [W-IT-FIN]. Also helpful would have been to experience validation after following an AI recommendation, or seeing that the combination between one's judgement and the recommendation was effective. "*Back test on the output and learn/evaluate the experience from there*" [W-IT-FIN], "*feedback on the results, self-learning*" [W-IT-FIN]. This came up more strongly with participants from the finance sector because of the nature of the results of their decision-making which is more 'easily' measurable (e.g., return of a loan). "*If there is a track record of successful cases for the use of an algorithm, then they would use it.*" [W-IT-FIN] Participants stressed the exponential element related to the use of proven effective recommendations, and how the more they would use and get positive feedback, the more they would use it. "*Time, and number of people (aka how much the AI has learned)*" [W-IT-FIN].

Credibility of the algorithm was listed even as one aspect that if strongly present and proven would have made people second guess their own beliefs, only second to what previously discussed with regards to organisational influence and the related factors, as organisational limitation of resources and therefore policies and selection. "*High demand of loans and low resources*" [W-IT-FIN].

### 6.2.4.2 Ecological validity: Selection of the characteristics in the experiment vs. real life

One of the main results of the discussion, and feedback to the study, is related to the ecological validity of the study. We collect and analyse here the participants' feedback on how they think the set-up of the experiment might have impacted their experience.

#### A. Choice of characteristics and their importance

In the discussion of the ecological validity of the characteristics, we discussed with the participants some of the preliminary results of the quantitative study. As shown in Figure 21, starting from original characteristics considered by the AI, participants proceed to add other relevant characteristics, comment on them, and order them by importance ("must have", "optional").

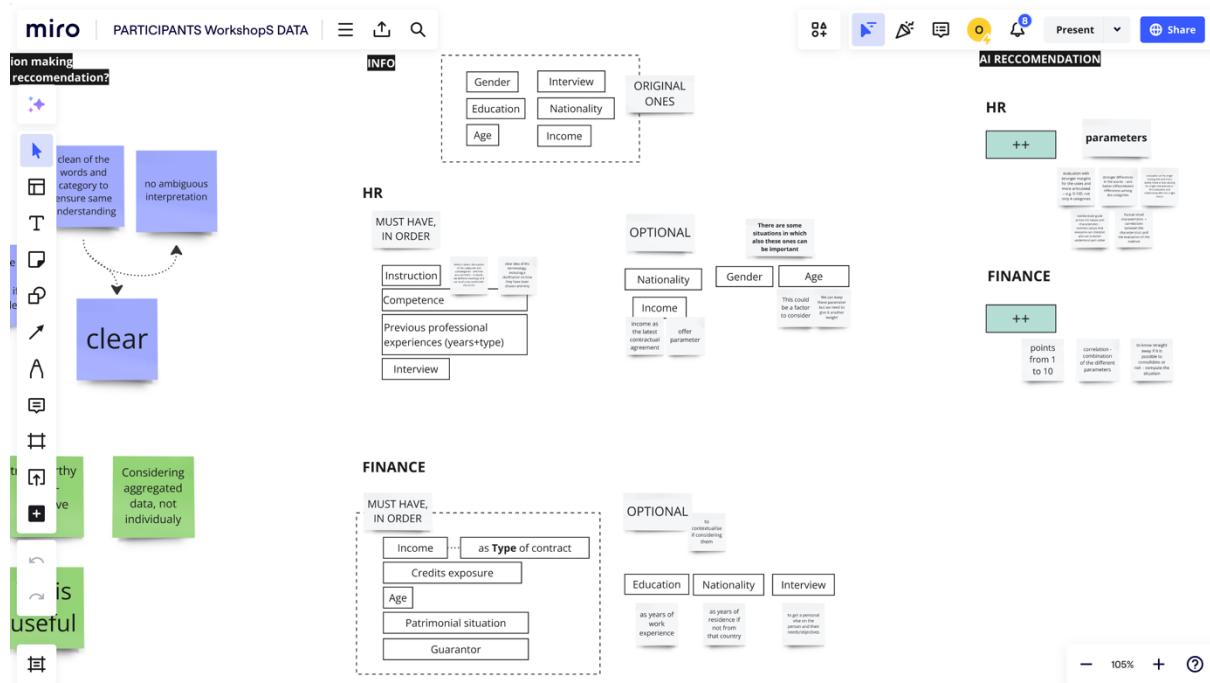
---

<sup>27</sup> This finding, whereby participants may be able to elicit and take account of specific types of information that are not accessible to the AI, is one of the main reasons evoked in (Balakrishnan, Ferreira, and Tong 2024) for allowing human interventions and override of AI systems.

The characteristics considered by the AI appeared not to be equally relevant for both sector specialists (e.g. income for HR or education for finance). Furthermore, participants expressed surprise at how “interview” didn't make it to the shortlist of the most important characteristics that influenced choice in the quantitative study, as it was considered quite important for the participants of the qualitative study, specifically in HR. “shocked interview is not here” [W-IT-HR] “The interview is really the moment in which you understand the candidate and their profile” [W-IT-HR]. The same type of remarks occurred for “Income” for Finance “Where is Income? Crazy it is missing” [W-IT-FIN].

Overall, during the discussion, “interview” was agreed to be the most relevant factor for HR, and “income” for Finance. Interviews were described as important to extract implicit elements, fundamental for the final decision. This important step for human decision-making, while highlighted as fundamentally a human task, it is also potentially a high-risk space for the development of possible biases.

Figure 21 Screenshot of the Miro Activity run during the workshop, in specific regarding the discussion on characteristics and ecological validity of the experiment.



## B. The AI recommendations format

During the discussion of the was the AI recommendation were given in the experiment, participants listed what they thought would have helped their understanding and increase their uptake.

“Evaluation with stronger margins for the votes and more articulated -- e.g. 0-100, not only 4 categories” [W-IT-HR], “stronger differences in the scores and better differentiated differences among the categories” [W-IT-HR], “number and short description” [W-IT-HR]

Participants didn't think the scoring was a sufficient recommendation/explanation as often confusing and open for interpretations. “Without explanations I do not feel well informed about the algorithm, so I follow my own perception more than the suggestion” [W-IT-HR]. “Clear idea of the terminology, including a clarification on how they have been chosen and why” [W-IT-HR].

Participants related the trustworthiness of the algorithms with its transparency and the access to explanations that they felt the experiment lacked. *"AI choices/recommendations need to be contextualised"* [W-IT-FIN], *"without justification there is a high chance of doubts, confusion and questioning - it becomes more of a disturbance than useful"* [W-IT-FIN], *"to contextualise the values and whys"* [W-IT-FIN], *"Perspective of who take the decision and the possible interests"* [W-IT-FIN].

### 6.2.5 Results from the workshop with experts

Figure 22 Workshop with experts during the group work discussion



Source: Own material

As a follow-up of the quantitative experimental study and the qualitative study with interviews and participants' workshops, we conducted a participatory speculative workshop with experts. More specifically, with this workshop we aimed to address the RQ4 of this study namely:

RQ4: How can we envision a fairer hybrid system of algorithm-supported human decision-making process in the scenarios under examination and in other real-life scenarios?

The expected outcome of this workshop was a set of ideas and opinions about algorithms fairness and biases in AI-supported human decision-making that would contribute to the interpretation of the results from the qualitative and quantitative studies but also as the starting point for the last part of the study which involves a workshop with policymakers working at the European Commission.

The participants discussed widely about algorithmic and human fairness tackling different aspects of it. Below we list the main topics that were discussed and then go on to analyse them:

1. Defining algorithmic and human fairness
2. Human-AI collaboration and how to operationalize fairness.
3. Ethical considerations and regulatory needs for human oversight
4. A futuristic perspective on the role of AI: Mutual checks
5. Sense of control and the maker-role
6. Proposed directions, responsibility, and context

### 6.2.5.1 Defining Elements of Fairness and Discrimination in AI, human, and hybrid systems

During the workshop, participants identified the need for a definition of fairness in the context of algorithmic and human decision-making. We explored the evolution of the concept from the past to what it is now and what it might be in the future. We used the medium of tarot cards, to ask people to symbolise this present or future ‘reading’. Participants envisioned fairer hybrid systems of the algorithm-supported human decision-making process as “*an internal moral compass of each and everyone, including policymakers. Looking into yourself through AI*”. [Expert workshop participant, “EWP”]

Although it was not possible to conclude with a specific definition of algorithmic and human fairness, participants’ discussions yielded aspects that should be considered when defining fairness. Below, we summarize the main points of discussion:

- a. Fairness as a dynamic process: One of the main outputs of the discussion was on shifting the concept of fairness from static to a rather dynamic process that needs to be practiced and exercised continuously. Participants discussed that a useful conceptualization of fairness would be its perception not as static concept but rather a dynamic process that needs to be practiced and exercised continuously. Fairness should be integrated into the entire process of developing AI systems, including data collection, training, and implementation. The responsibility for ensuring fairness lies with the makers of the systems, and the focus should be on establishing fair processes.
- b. Fairness as a systemic and contextual construct: The experiment started with a specific focus on discrimination against individual people, while in the discussion and contextualisation of the work, more systemic aspects of discrimination were raised and highlighted. Fairness is contextual and depends on the specific circumstances and characteristics involved. Achieving fairness in AI requires a multidisciplinary approach, involving both social and technical considerations, and building collaborative fairness requires awareness of when the system is failing and when humans are failing for the system. Reflections are particularly necessary and relevant for a European organization, as tackling discrimination against individuals, is not enough, if the system does not change. One of the conclusions that were discussed during the workshop is that fairness is contextual and depends on the specific circumstances and characteristics involved. It is important to go beyond our current biases and consider various measures of fairness, such as equal treatment and equal opportunity.
- c. Fairness and protected characteristics: During the workshop, participants formulated questions around the role of the inclusion or exclusion of information about people’s protected characteristics among variables shown to deciders. Participants wondered: “*Does excluding sensitive attributes guarantee fairness?*” [EWP] and stated “*fairness through unawareness is not a popular attempt at fairness*” [EWP]. Participants discussed the definition of fairness by excluding protected characteristics and whether removing them from a dataset would guarantee fairness. Along these lines, a wider observation, focusing on the consequences and context of defining fairness in this way highlights that intentionally ignoring certain characteristics or factors in decision-making processes may not be an effective approach to achieving fairness. Other comments focused instead on the type of characteristics that have been defined as protected:
  - “*Age/education/income could be also articulated as potentially sensitive metrics?*” [EWP]
  - “*Are all decisions made by a discriminatory system inherently unfair or if some decisions could still be fair due to the influence of other characteristics?*” [EWP]
  - Other observations on the characteristics and fairness focused on the idea that by isolating them, and not looking at them as a system the possibility of them being collectively telling

another layer of interpretation or that by contextualising them in a fair or unfair society, these could be lost.

- d. AI-Human mutual transparency into fairness: Achieving fairness in AI requires a multidisciplinary approach, involving both social and technical considerations. Building collaborative fairness requires awareness of when the system is failing and when humans are failing for the system. Transparency plays a crucial role in achieving this awareness.

#### 6.2.5.2 Human-AI collaboration and how to operationalize fairness.

For this study, we use the term “collaboration” between AI and humans interchangeably with the term “interaction”. While the terms might imply the attribution of agency in the AI, this is not our goal, and we used the term according to current conventions in the field of human-computer interaction. Participants who discussed the scenarios of the study suggested that the collaboration between humans and AI agents should be the objective, where both parties support and assist each other. Furthermore, they highlighted how building collaborative fairness requires awareness, meaning that achieving fairness in AI systems involves being aware of when the system or the human is failing, which can be only partially achieved through transparency.

#### 6.2.5.3 Ethical considerations and regulatory needs for human oversight

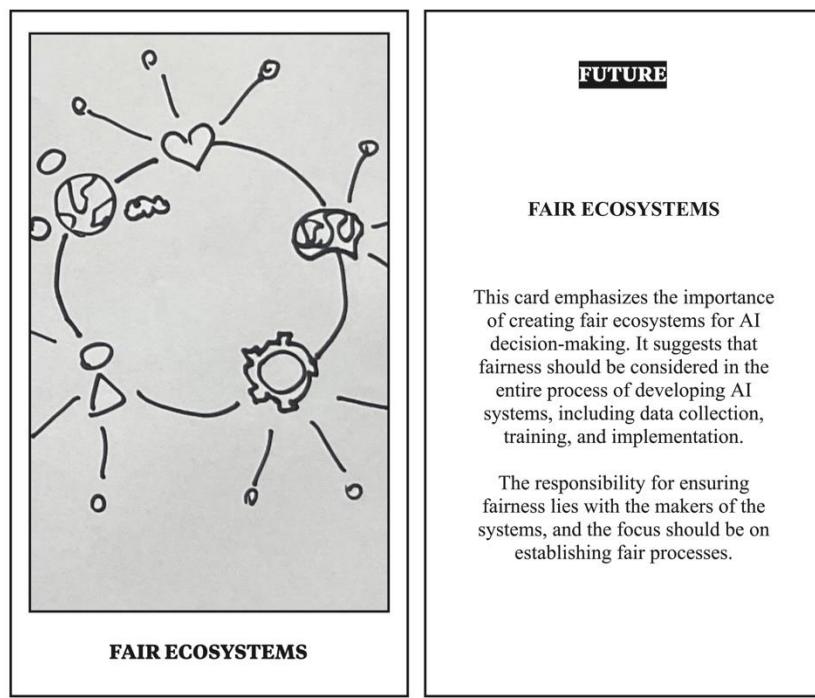
- a. Ethically informed feedback: Participants discussed the need to design and develop AI systems that embed *“ethical hiring practices”*. In this way, the system would support non-discriminatory and fair treatment of all candidates. Ideally, the AI supporting system should be regularly updated to comply with evolving legal standards.
- b. Mechanisms of standardized control and feedback loop: Even in cases of ethically aligned systems, participants expressed their belief in the importance of “detection of human/AI agent suspicious behaviour”. Participants described feedback loops (from both AI and human suggestions) as a tool for the improvement of decision-making processes. In this process, humans and AI systems can be informed about the possible limitations, successes, and overall results of certain decisions that would (positively or negatively reinforce the (human or artificial) agent. This mechanism would focus on the process and what the agent could learn from this feedback. For recruiters, the assistant provides a detailed analysis of each candidate, highlighting their suitability for the role. It also offers insights into market trends, salary expectations, and the competitive landscape to inform recruitment strategies.
- c. Meaningful analysis and transparency towards the candidate: In addition, this mechanism should be accessible to the candidate. A possible scenario described by the participants included an instance in the *“post-interview, the assistant collects feedback from both parties, using this data to refine future matches and provide constructive feedback to candidates”*. However, as mentioned by participants, companies should use this kind of mechanism with caution since *“Feedback loops should not exploit underrepresented individuals to make your system better”*.

#### 6.2.5.4 A futuristic perspective on the role of AI: Mutual checks

Throughout the workshop, participants discussed the challenges in human-AI interaction in decision-making and identified possible improvements. One of the prevalent aspects was the complementarity between humans and AI and most importantly the possibility of mutual checks. Participants mentioned that human and AI interaction is crucial for decision-making, with *“AI serving as a mirror for human biases and providing support and questioning of actions”*. Iterations between humans and AI could facilitate a process that would involve challenging each other rather than deciding for one another. During the discussions, participants posed the following questions:

- "How to post-evaluate a decision made by AI within a specific organizational context?" the experts highlighted the importance of providing feedback to the end result of a suggestions made by the AI, looking at organisational learnings and contextual practicalities.
- "How can we support deciders to be aware of their own biases?" in this mutual relationship the level of awareness of the overseer plays a role in recognising of their biases, and their preparedness to question the proposed suggestion.
- "In the age of complexity, could we see AI as a 'trust proxy?'" another discussion point revolves around the idea of 'trust by proxy' and how trusting the system because "someone/something that I trust has trust in you". The relationship of proximity and trust could influence the mutual check and decisions, especially in a complex and intertwined context.
- "Should Human & AIs challenge each other (frictions) rather than decide for one another?" on this idea of relationship, the elements of 'friction' and 'challenging' each other came up as a key element of the interaction, trying to understand how the 'mutuality' through the challenge could help establish the ultimate balance.

Figure 23 One of the tarot cards that represent the “future” definition of fairness in the AI+Human decision-making process.



Source: Own material

#### 6.2.5.5 Sense of control and the maker-role

As the main topic of the whole study was to better understand what human oversight means in AI-supported decision-making, participants often discussed one of the basic requirements for it which is human control. Throughout the workshop, control took various forms and was mentioned as a characteristic of varied intensity, starting with control over the training data of an algorithm and ending with the human taking the role of "maker". Below, we some of the relevant highlights:

- a. Sense of control: One of the basic characteristics of human oversight that was prevalent in the discussion was the need for a sense of control of different stakeholders, with a focus on recruiters. Participants distinguished between the processes where humans indeed have control in the decision-making process and those cases where they perceived as being in control. "*Humans want to be in control - some sort of perceived control*" [EWP].
- b. Shared checked dataset: The sense of control was not only discussed regarding the recruiters but also regarding other stakeholders. For example, participants discussed whether companies would prefer to use their own data and models or external ones. In any case, the participants indicated that a distributed system might trigger more shared responsibility. Systems that are shared will be easier to test and audit and as such this might be one of the ways to attain fairer systems while on the other hand, relying on their own data or models to build a proprietary system will lead to competitive advantages. Participants discussed how data and models should be tested against a fairness 'reference system', a sort of 'sample' to test and compare systems to: "*My data to be verified against fairness test sample*" [EWP]. Overall, with fairer data sets and models smaller lenders will benefit too, which refers to greater societal fairness.
- c. **The AI "co-pilot":** Although it was not among the intentions of the research team to impose any vocabulary regarding the characterization of the interaction between humans and AI, participants identified one of the characteristics of AI-enabled decision-making, its "collaborative" nature. However, they also highlighted the importance for the person on the hook for the decision to be AI-literate and needs to be enabled to have a discussion, revise, and challenge/be challenged. For example, participants discussed how the professional should be able to defend and explain themselves and the behaviour of AI to a third party; regarding the actual performance, the professional should be able and open to serve as an advisor to AI "detection of human advisor behaviours". Human oversight was considered contextual by the participants of the workshop.
- d. **The maker's role:** In addition to the AI literacy of the users, participants highlighted the importance of AI Ethics Literacy of the designers and developers. Good knowledge of the potential risks and the needs of the users of AI is necessary for the designers of AI-based decision support systems. Participants framed this discussion around trust. Some of the questions they posed were "*Did somebody I trust design it?*", "*Was I involved in the design?*", "*Does it reflect my values and vision?*".
- e. Fairness and Trust as a process: In the same direction, participants highlighted the importance of understanding that "human-human interaction is key for the deciders". This was connected to the feeling of trust in AI systems. Building trust in AI systems requires transparency in how recommendations and decisions are made, as well as awareness of biases and the involvement of users in the design process. Fairness and trust were seen as a process that involves interactions and feedback loops. With these processes, human deciders are familiar when they involve interactions with humans, but it requires new strategies for them to develop fairness and trust in AI support systems.
- f. Community in the loop and challenges in shared responsibility: Lastly, during the workshop, participants often discussed that fairness should be perceived as a contextual and collective action. Fairness exists in the context of a certain community, with certain norms and for this reason, human oversight of AI systems can take a collective form and be a collective action. A "Community in the loop approach", so having members of the community in one way or

another to check and being involved in the decision making ‘system’ was one of the proposals of the experts.

#### 6.2.5.6 Proposed directions, responsibility, and context

With one of the activities of the workshop, we asked the experts to write a fictional letter to the policymakers of the European Commission to include their proposals for future directions. Human oversight was discussed as a concept, as an experiment filter, and as a general process to follow. Experts emphasize the need to prioritize values over technologies, invest in trust as a key aspect of technology, and build collaborative communities. The letters also highlight the importance of addressing societal issues alongside AI development, involving society in decision-making, and creating frameworks for a fair society in the context of AI. Below we list the points experts included in this exercise:

- a. Center Values and Communities in Techno-social Collaboration. This means starting with the needs of communities rather than focusing solely on technical capabilities. Building new imaginaries of techno-social collaboration that prioritize protopian changes and center lived experiences can lead to more inclusive and fair AI development.
- b. Build Stronger Frameworks for a Fair Society in the Context of AI. This would involve addressing societal issues such as discrimination, crime, under-education, and digital exclusion, which are not solved by AI alone. It is important to develop a socio-technical supporting ecosystem that surrounds the use and abuse of AI, involving society in the support and defence of those affected by AI.
- c. Develop an Interactive Continuous Process for Determining Fairness in AI systems. This would involve creating mechanisms to assess and monitor the fairness of AI algorithms and models and the impact on society and individuals.
- d. Establish EU Policy Principles, Ethics, and Values for AI. By prioritizing these aspects over purely technical considerations, we can ensure that AI development aligns with its desired outcomes. This includes setting standards, developing better AI evaluation systems, and creating taxonomies for innovation.
- e. Invest in Trust and Slow AI Growth. Europe should invest intellectually, legislatively, and financially in trust as a main axiom for technology. Instead of solely focusing on growth as a metric, Europe should prioritize building trust in AI systems. This involves taking a cautious approach to AI growth and considering the impact on society and individuals.
- f. Include Fairness in EU Calls for Funding Research to boost AI research and development. This would involve incorporating existing definitions of fairness and considering all dimensions of fairness.

## 7 Discussion

With this study, we aimed to deconstruct and explore how fairness is affected in the interplay between AI decision support systems and human decision-making and how human oversight might improve (or not) fairness in the final decision. The results of this research can inform the design and policy recommendations regarding human oversight for AI-supported decision-making systems.

We addressed our research questions with a mixed method approach which integrated:

- a large-scale experimental study with professionals in the field of human resources and banking; and
- a follow-up qualitative approach with which we combined inputs from
  - interviews and focus-groups from a subset of the participants.
  - workshops with a multidisciplinary group of experts.
  - Workshops a group of policymakers working at the European Commission.

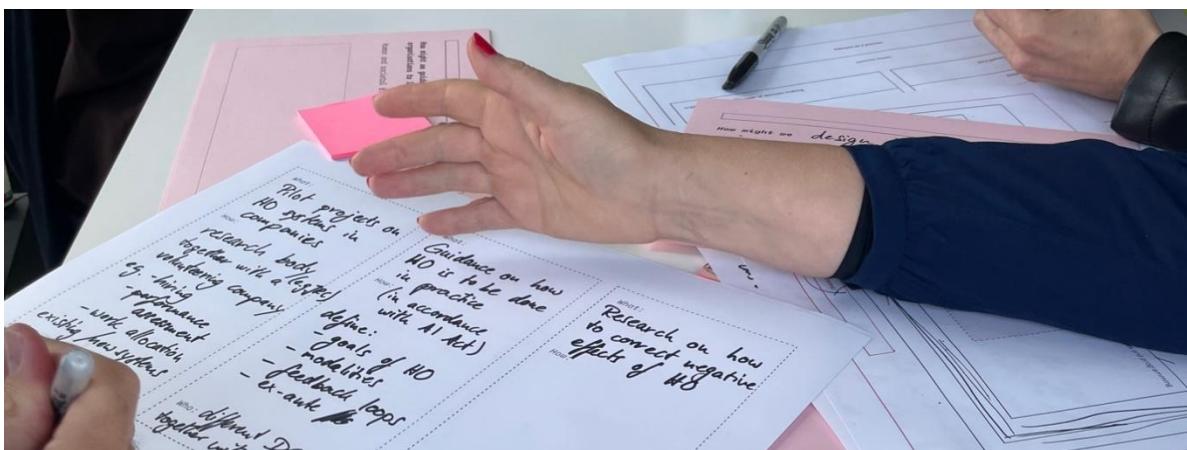
This approach provides a holistic view into the problem of fairness in AI-supported human decision-making and of the role of human oversight.

We followed established inductive and deductive methods, iterative approaches and we took a multi-disciplinary approach, and we considered the findings from the quantitative and qualitative study as complementary. The large-scale experimental study indicates generalizable trends in the specific population, while the qualitative results allow for in-depth explanations and contextualization of the findings while indicate the dynamics of interactions between humans, algorithms, organizations and contextual information.

Since the scope of this work was not only to provide scientific evidence about AI-supported human decision-making but to explore policy relevance and eventually provide policy guidance in relation to human oversight, the last part of our study included a workshop with policymakers working at the European Commission.

We organized an interactive participatory design session with a diverse group of policymakers which was held in Brussels in June 2024 (Figure 24). The scope of the session was to discuss the results of the study and to reflect on actionable next steps for policymakers.

Figure 24 Workshop with policy-makers in the EU Policy Lab, discussing opportunities and needs in regards to Human Oversight



Source: Own material

Policy makers worked in small groups to brainstorm practical implications for implementation of existing policies. Each group focused on one of the opportunities outlined, from the overseer, the human and AI complementarity and the final decision outcome. This activity aimed to bridge the gap between theoretical insights and practical implementation strategies. Policy makers were tasked with extracting actionable insights from the workshop discussions and formulating recommendations that could be adopted by the European Commission and other relevant bodies.

Groups presented their recommendations, which included proposals for new regulatory guidelines, initiatives for stakeholder engagement and training, and strategies for monitoring and evaluating AI systems. The session concluded with a discussion on how to advocate for these recommendations within the European policy landscape, identifying key stakeholders and potential partnerships for effective implementation.

Whilst the findings section of this report presents separately the quantitative and qualitative study results, with the discussion emerge the two to analyse the combined results and to propose next steps based on the integration of the two approaches.

As a final step, we integrate the quantitative and qualitative results together with the results from the policymakers' workshop to propose opportunities for actionable steps. Some of this ideation of solutions was already explored during the workshop with professionals and is explored as a set of possible interventions in the results section.

Table 23 Main findings and opportunities to improve human oversight.

	Overseers	Human + AI	Decisions		
Finding s	Overseers go along with AI discrimination if fulfilling the norms and objectives of the organisation.	Overseers override AI decisions in part to fit their own discriminatory preferences.	Overseers see their value in being able to assess a candidate's specific situation. They think they can better assess "soft" attributes of candidates	Overseers underlined their lack of experience with AI systems. They need feedback on whether the AI-supported decisions are correct.	Human oversight can introduce biases in the outcome of AI-supported decisions.
	↓	↓	↓	↓	↓
Opportunities to explore	How might we guide organisational norms to obtain less discriminatory outcomes?	How might we oversee and review the decisions to override to detect potential biases and improve the AI system?	How might we enable critical and complementary AI-Human decision making so overriding is based on factors that can be judged only by humans?	How might we enable humans to receive and provide regular feedback from and to the AI-supported system?	How might we monitor the outcomes from the use of AI so the AI system is fair ex-post (in terms of outcomes)?
Domain to address	Human and organisation al biases	Oversight of the overriding	Mutual checks and reinforcement	Outcome feedback and reinforcement learning	Outcome monitoring

Source: Own analysis

Table 23. summarizes our critical reflections on the main findings of this study by combining findings from the quantitative and qualitative study and the workshop with the policymakers.

The rest of this chapter discusses the integrated findings with the corresponding opportunities for next steps which can function as a policy recommendation for the framing of human oversight in AI-supported human decision-making.

To conclude and give space to further work, we intentionally conclude framing from the study insights, the opportunities, under the format of 'How might' and a generic 'we' to enable each actor to pick up and define their 'we' and needs based on the circumstances, and scale they are dealing with. We are providing below with an analysis and grounding of the insights into the research outcomes and providing the space for further exploration.

## 7.1 Human and organisation biases (Overseer)

Main finding 1: Overseers go along with AI discrimination if fulfilling the norms and objectives of the organisation.

Opportunity 1: How might we guide organisational norms to obtain less discriminatory outcomes?

We observed that individual decision-making was influenced by organizational, contextual, and systemic predispositions, often embedded in decision-support systems. For example, our participants explained that if their company or organization instructed them in a certain (biased) way, they would follow these instructions, even if they discerned that an AI recommendation favours people with certain characteristics. A consequence of the unwillingness of overseers to challenge organisational norms is that they are likely to adhere to AI-supported discrimination if presented to them as fulfilling norms and objectives of the organisation. Indeed, in our experiment, overseers went along with discriminatory AI as often as they went along with fair AI.

A recent survey on human oversight policies based on empirical research highlighted like us the need for institutional oversight to complement human oversight of algorithmic decisions (Green 2022). However, our study also indicates that institutional oversight could also fail if organisational norms -- the unwritten rules and shared expectations that guide behaviour within an organization -- lead to biases. Organisations that lack diversity and do not examine how their norms and practices may lead to unequal opportunities both for their members and for their clients, are unlikely to be able to put in place oversight systems that promote fairness.

This analysis shows that it's important to connect discussions about fairness in AI with a wider range of existing EU policies. These policies and regulations are not just specific to AI but also include broader measures that the EU has put in place to fight discrimination, ensure equal treatment, and support social inclusion. Therefore, when we consider how AI should be regulated and implemented, we must take into account and integrate the full spectrum of EU initiatives that aim to protect fundamental values.

## 7.2 Oversight of the overriding (Overseer)

Main findings 2: Overseers override AI decisions in part to fit their discriminatory preferences.

Opportunity 2: How might we oversee and review decisions to override to detect potential biases and improve AI systems?

Our empirical analysis, in line with the existing literature, indicates that in the specific scenarios of our study, human biases influenced choice even when a fair AI was provided to support decisions, and thus could potentially have helped deciders not to let their own biases influence choice. Furthermore, human oversight did not prevent bias present in the generic AI to express itself in terms of final choice of applicants by the deciders. Those issues should be taken into account and a more elaborate approach of human oversight should thus be considered. Indeed, many participants in our study showed biases themselves based on their previous experiences or existing societal biases. Those were reflected into selective algorithm aversion.

The necessity to review decisions to override goes beyond measures that are usually included as part of algorithmic auditing (Sandvig et al. 2014). Our research points out that we must go beyond examining the decisions made by AI systems to identify and mitigate biases, ensure fairness, and improve the overall reliability of these systems. Having a Human-in-the-Loop (HITL) who can give real-time feedback and adjust outcomes is not enough and can be counter-productive (Green 2022).

Bias detection tools must be adapted to not only detect biases in AI systems, but also biases in the oversight of AI systems. Regular reviews and updates of AI systems should therefore include not only re-evaluating data sources and algorithms, but also decision-making processes and their outcomes. This requirement should thus be part of ethical guidelines and legal frameworks that guide the development and deployment of AI systems.

## 7.3 Mutual checks (Human + AI)

Main finding 3: Overseers see their value in being able to assess a candidate's specific situation. They think they can better assess "soft" attributes of candidates.

Opportunity 3: How might we enable critical and complementary AI-Human decision making so overriding is based on factors that can be judged only by humans?

In this study, we intentionally simulated situations where professionals should take the final decision about a candidate by following or not the recommendations of a fair or unfair AI system; however, during interviews and workshops with professionals, we observed that professionals' decision to follow the AI recommendation or not did not only depend on whether the AI was fair or not but on the type of AI contribution and the context of operationalization. Professionals would trust more an AI recommendation that deals with less complex tasks, such as data organization, analysis, and simple processing, while for more complex tasks, *i.e.* assessing a candidate's "soft skills" or situational, contextual, and implicit information, they trusted AI less. As such they questioned the "interview" score of our AI.

Relevant literature indicates that forming human-AI teams, in which the AI system augments one or more humans by recommending decisions, while people retain agency and have accountability for the final decisions is a viable but challenging scenario (Bansal et al. 2021; Paul Hemmer et al. 2024). One key challenge in AI-assisted decision-making is whether the human-AI team can achieve

complementary performance, i.e., the collaborative decision outcome outperforming human or AI alone (Bansal et al. 2019). This becomes even more challenging when considering possible algorithmic and human biases, user's domain expertise, mental models of an AI system and trust in recommendations which can impact the success of Human- AI teams (Inkpen et al. 2023). Our participants envisioned a possible scenario for a complementary human-AI interaction in which both humans and algorithms would perform mutual checks and reflection; especially in cases like the ones in the present study, where societal and individual biases intertwine and may lead to unfair decisions from both the AI systems and humans.

Several approaches to involve humans in AI decision making have been proposed. They define the situations in which a human should be called on to review AI decisions and the conditions under which such a review is likely to be productive. Situations warranting intervention include when the AI's confidence is low, when the situation is ambiguous or sensitive (Attenberg, Ipeirotis, and Provost 2011), when the decision-making process is complex (Parasuraman, Sheridan, and Wickens 2000), or when the decisions involve ethical considerations that go beyond the AI's programmed capabilities (Greene et al. 2016). This is the topic of Human Centered AI (Shneiderman 2020). Conditions under which review is likely to improve outcomes are when the AI's decision process (reasoning) is (made) understandable to humans; this is the topic of Explainable AI (Lipton 2016).

Our findings suggest that humans should be given instructions and when, why, and how to override decisions by AI. Their decisions to override should be reviewed both to improve the AI systems but also to detect possible biases in overriding that would favour some groups at the expense of others. In the same way as AI should be explainable and justify its decisions, decisions to override AI should be documented, monitored, and require explanation. The identity and characteristics of the "Human in the Loop" should also be taken into account to better understand their decisions to override.

## 7.4 Outcome feedback and reinforcement learning (Human + AI)

Main finding 4: Overseers need feedback on whether the AI-supported decisions are correct.

Opportunity 4: How might we enable humans to receive and provide regular feedback from and to the AI-supported system?

Research in cognitive science indicates the core role of past experiences in human decision-making process (Aarts, Verplanken, and Van Knippenberg 1998). Learning from experience rewires human brain so that it can categorise the objects and concepts we are looking at and respond appropriately to them in any context. In an analogous way, ML algorithms allow us to model and predict big data behaviours based on historical data (Christiano et al. 2017).

In addition to human past experiences and memories and algorithmic historical data, reinforcement learning that includes future goals is an effective method in decision-making in both humans and algorithms. By this we mean that reinforcement learning should take into account the desired outcomes that a person or an algorithm aims to achieve at some point ahead in time. These "future goals" help guide the decision-making process by providing a target or objective to work towards, in our case, some notion of fairness in decision-making. In this way, reinforcement learning is not limited to only replicating past, possibly discriminatory decisions.<sup>28</sup>

<sup>28</sup> Goal conditioned reinforcement learning is a method of learning that does not rely on pre-set rewards and punishments as in standard reinforcement learning, but rather guides learning by its ability to get close to a goal (Liu, Zhu, and Zhang 2022). This replicates some of the ability of humans to act based on the attainment of desired objectives independently of knowing the reward function they are facing in a situation (Molinaro and Collins 2023; Veksler, Gray, and Schoelles 2013)

In complex multifaceted situations, however, reinforcement learning is a challenging problem and the necessary feedback is often lacking because (i) variability of the environment degrades the reliability of the decision or recommendation (ii) outcomes are delayed and not directly correlated with a particular action or characteristic of a candidate, and (iii) there is no feedback about what the outcome would have been if a different decision would have been made (Tversky and Kahneman 1986). In this study, understanding and providing useful feedback requires actual time, as it can take a while before knowing if a decision taken for a new hire was successful or not, or whether and how a loan is paid back.

Effective human-AI teaming and human oversight in decision-making thus remain challenging when the future goals that are communicated to the hybrid team involve ethical and moral elements such as fairness or accountability (Charisi et al. 2017). Accordingly, the participants of this study indicated that reinforcement learning should combine human and algorithmic feedback in a systemic way and that explanations of reasoning from both sides (AI and human) can improve the final decisions (Schmude et al. 2023).

## 7.5 Outcome monitoring and alignment (Decisions)

Main finding 5: Human oversight can introduce biases in the outcome of AI-supported decisions.

Opportunity 5: How might we monitor the outcomes from the use of AI so the AI system is fair ex-post (in terms of outcomes)?

This study shows that ex-post human oversight can lead to the introduction of biases in the outcome of AI-supported decisions. Not only does ex-post oversight by humans not prevent harmful consequences of the use of discriminatory AI, but it can in fact introduce bias even when using fair AI. The participants of the present study indicated that a mere presentation of candidates' scores, despite a form of explanation for respective AI recommendations, was not enough for the professionals to trust the AI recommendation. As such, reinforcement with the outcomes of the final decision were proposed as *dynamic, contextual, and continual processes* that monitors the outcome of a decision in the short and longer term and within various contexts. Interestingly, one of the views for a hybrid reinforcement approach is the regular feedback by human decision-makers which can be used not only to improve a current decision but also to contribute to fairer AI. Outcomes from the use of AI should be monitored. AI system should not only be programmed to be fair ex-ante (in theory), but also fair ex-post (in terms of outcomes). Similar approaches are currently being explored by teams that work on *AI alignment* (Gabriel 2020). They explore technical solutions on how AI systems can be formed and shaped by human values. Overseers saw their value in being able to assess a candidate's specific situation more finely, or evaluate dimensions the AI may not otherwise consider, e.g. the soft skills. Human criteria may change over time and a static algorithmic preference assumption may undermine the soundness of recommendations leading to implicitly reward AI systems for influencing user preferences in ways users may not truly want (Carroll et al. 2024). As such, monitoring decision-making outcomes and establishing mechanisms that consider the dynamics and the current limitations of human-AI interactions in decision-making seems needed. Auditing algorithm should not rely only on "black box" access, i.e. interpreting the outcomes of algorithms to check they are fair, reliable and accountable. One should also be able to obtain "white box" access (inner working)- and "outside-the-box" access, i.e. be able to examine the algorithm's development process (Casper et al. 2024). This would allow to better identify differences between what the algorithm was programmed to attain (the goals of the developers), and what the users actually want it to do. In turn, this would allow users to give feedback more efficiently to developers and influence development towards their own goals.

## 8 Conclusion

Our study underlines the difficulties in achieving fairness in AI-assisted decision-making. Most people intend to make fair decisions, but doing so is not easy. We tend to think we are in control of our own decisions, but they are affected by societal norms, laws, rules, past experiences, and social practices that we often are not aware of or do not question. Our study showed that human oversight of AI was not enough to prevent discriminatory outcomes from the use of AI, and could even worsen them. Human overseers were prone to algorithm aversion, especially when AI recommendations conflicted with their preferences. Yet, they readily accepted them when AI guidance aligned with their own biases.

We do acknowledge that human oversight is crucial in managing AI systems. However, our study showed that human overseers brought their own biases, predispositions, values, and past experiences. This affected the outcome of joint AI-Human decision making, both for good and for bad. Decision-makers were not fully capable of detecting and rejecting biased AI recommendations. Even though we made the bias of the AI transparent to the overseer, this was insufficient to deter or moderate its use.

Our study leads to the suggestion that it is not enough to program AI that respects fairness norms. We must also put in place *oversight systems* that prevent human bias from influencing the outcomes of the AI advisory relationship. Efforts to ensure non-discriminatory outcomes should therefore focus on the development of better *societal guardrails* to guide individual decision-making. Those guardrails can both be necessary and effective. Indeed, in the same way that society can generate biases in individuals, it can also provide the tools to correct them (Gasser and Mayer-Schönberger 2024).

Those guardrails could include bias training for overseers, guidelines for when to override AI, and regular audits of AI-influenced decisions. A range of measures could also ensure a meaningful role for the human overseer. AI DSS should allow overseers to take account for additional decision-relevant information that are not taken account of by the AI. Overseers should also be able to influence development by raising issues with AI developers. This would promote reliance on the AI system while ensuring it adapts based on human feedback.

Further research should explore how these oversight mechanisms can be implemented in various contexts and their effectiveness in different scenarios. We also recommend that AI developers and policymakers collaborate closely to translate these findings into practical regulations and standards.

In partnership with technical communities that are dedicated to aligning AI systems with human values, our systemic and procedural approach to oversight goes beyond the individual level to encompass the entire decision-making ecosystem. This approach ensures that AI systems are not only technically sound but also ethical and socially responsible.

Our study underscores the importance of addressing AI decision-making from a systemic fairness perspective, and of considering both the human and the AI to understand their final combined decisions. Such an approach promises to align AI with societal values and ethical norms, fostering trust and ensuring that AI systems function effectively within their intended contexts.

## References

- Aarts, Henk, Bas Verplanken, and Ad Van Knippenberg. 1998. "Predicting Behavior from Actions in the Past: Repeated Decision Making or a Matter of Habit?" *Journal of Applied Social Psychology* 28 (15): 1355–74.
- Alon-Barkat, Saar, and Madalina Busuioc. 2023. "Human–AI Interactions in Public Sector Decision Making: 'Automation Bias' and 'Selective Adherence' to Algorithmic Advice." *Journal of Public Administration Research and Theory* 33 (1): 153–69. <https://doi.org/10.1093/jopart/muac007>.
- Amnesty International. 2021. "Xenophobic Machines: Discrimination through Unregulated Use of Algorithms in the Dutch Childcare Benefits Scandal." Amnesty International. October 25, 2021. <https://www.amnesty.org/en/documents/eur35/4686/2021/en/>.
- Amnesty International and Access Now. 2018. "The Toronto Declaration." 2018. <https://www.torontodeclaration.org/declaration-text/english/>.
- Attenberg, Josh, Panagiotis G. Ipeirotis, and Foster Provost. 2011. "Beat the Machine: Challenging Workers to Find the Unknown Unknowns." *Human Computation* 1 (1): 160–75.
- Auger, James. 2013. "Speculative Design: Crafting the Speculation." *Digital Creativity* 24 (1): 11–35.
- Avery, Mallory, Andreas Leibbrandt, and Joseph Vecci. 2023. "Does Artificial Intelligence Help or Hurt Gender Diversity? Evidence from Two Field Experiments on Recruitment in Tech." Monash Economics Working Papers 2023–09. Monash University, Department of Economics. <https://ideas.repec.org/p/mos/moswps/2023-09.html>.
- Balakrishnan, Maya, Kris Ferreira, and Jordan Tong. 2024. "Human–Algorithm Collaboration with Private Information: Naïve Advice Weighting Behavior and Mitigation." SSRN Scholarly Paper. Rochester, NY. <https://doi.org/10.2139/ssrn.4298669>.
- Bansal, Gagan, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. "Beyond Accuracy: The Role of Mental Models in Human–AI Team Performance." In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7:2–11. AAAI.
- Bansal, Gagan, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. 2021. "Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance." *arXiv:2006.14779 [Cs]*, January. <http://arxiv.org/abs/2006.14779>.
- Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.
- Bellamy, Rachel K. E., Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, et al. 2018. "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias." *arXiv*. <http://arxiv.org/abs/1810.01943>.
- Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. "Trust, Reciprocity, and Social History." *Games and Economic Behavior* 10 (1): 122–42. <https://doi.org/10.1006/game.1995.1027>.
- Carroll, Matthew, David Foote, Abhinav Siththaranjan, Stuart Russell, and Anca Dragan. 2024. "AI Alignment with Changing and Influenceable Reward Functions." *arXiv Preprint arXiv:2405.17713*.
- Casper, Sarah, Charles Ezell, Caroline Siegmann, Noah Kolt, Tracey L Curtis, Brian Bucknall, and others. 2024. "Black-Box Access Is Insufficient for Rigorous AI Audits." In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2254–72.
- Celi, Leo Anthony, Jacqueline Cellini, Marie-Laure Charpignon, Edward Christopher Dee, Franck Dernoncourt, Rene Eber, William Greig Mitchell, et al. 2022. "Sources of Bias in Artificial Intelligence That Perpetuate Healthcare Disparities—A Global Review." *PLOS Digital Health* 1 (3): e0000022. <https://doi.org/10.1371/journal.pdig.0000022>.
- Charisi, Vicky, Louise Dennis, Michael Fisher, Raul Lieck, Andreas Matthias, Marija Slavkovik, and others. 2017. "Towards Moral Autonomous Systems." *arXiv Preprint arXiv:1703.04741*.

- Charness, Gary, Uri Gneezy, and Austin Henderson. 2018. "Experimental Methods: Measuring Effort in Economics Experiments." *Journal of Economic Behavior & Organization* 149 (May): 74–87. <https://doi.org/10.1016/j.jebo.2018.02.024>.
- Charness, Gary, Anya Samek, and Jeroen van de Ven. 2022. "What Is Considered Deception in Experimental Economics?" *Experimental Economics* 25 (2): 385–412. <https://doi.org/10.1007/s10683-021-09726-7>.
- Charter of Fundamental Rights of the European Union*. 2012. OJ C. Vol. 326. [http://data.europa.eu/eli/treaty/char\\_2012/oj/eng](http://data.europa.eu/eli/treaty/char_2012/oj/eng).
- Christiano, Paul F, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. "Deep Reinforcement Learning from Human Preferences." In *Advances in Neural Information Processing Systems*. Vol. 30.
- Chugunova, Marina, and Wolfgang J. Luhan. 2022. "Ruled by Robots: Preference for Algorithmic Decision Makers and Perceptions of Their Choices." SSRN Scholarly Paper. Rochester, NY. <https://doi.org/10.2139/ssrn.4058066>.
- Creswell, John W. 2021. *A Concise Introduction to Mixed Methods Research*. SAGE Publications. <https://uk.sagepub.com/en-gb/eur/a-concise-introduction-to-mixed-methods-research/book266037>.
- Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err." *Journal of Experimental Psychology: General* 144 (1): 114–26. <https://doi.org/10.1037/xge0000033>.
- Eurostat. 2024. "Digitalisation in Europe." 2024. <https://ec.europa.eu/eurostat/web/interactive-publications/digitalisation-2024#technology-uptake-in-businesses>.
- Ferrara, Emilio. 2024. "Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies." *Sci* 6 (1): 3. <https://doi.org/10.3390/sci6010003>.
- Fetters, Michael D, Leslie A Curry, and John W Creswell. 2013. "Achieving Integration in Mixed Methods Designs—Principles and Practices." *Health Services Research* 48 (6 Pt 2): 2134–56. <https://doi.org/10.1111/1475-6773.12117>.
- Flick, Uwe. 2013. *The SAGE Handbook of Qualitative Data Analysis*. SAGE.
- Forlano, Laura, and Anijo Mathew. 2014. "From Design Fiction to Design Friction: Speculative and Participatory Design of Values-Embedded Urban Technology." *Journal of Urban Technology*, October. <https://www.tandfonline.com/doi/abs/10.1080/10630732.2014.971525>.
- France Info. 2022. "La Caisse des allocations familiales utilise un algorithme pour détecter les allocataires 'à risque'." Franceinfo. December 9, 2022. [https://www.francetvinfo.fr/economie/emploi/carriere/entreprendre/aides/enquete-la-caisse-des-allocations-familiales-utilise-un-algorithme-pour-detecter-les-allocataires-a-risque\\_5532651.html](https://www.francetvinfo.fr/economie/emploi/carriere/entreprendre/aides/enquete-la-caisse-des-allocations-familiales-utilise-un-algorithme-pour-detecter-les-allocataires-a-risque_5532651.html).
- Gabriel, Jason. 2020. "Artificial Intelligence, Values, and Alignment." *Minds and Machines* 30 (3): 411–37.
- Gasser, Urs, and Viktor Mayer-Schönberger. 2024. *Guardrails: Guiding Human Decisions in the Age of AI*. Princeton University Press. <https://press.princeton.edu/books/hardcover/9780691150680/guardrails>.
- Gerdon, Friederike, Robert L Bach, Christoph Kern, and Frauke Kreuter. 2022. "Social Impacts of Algorithmic Decision-Making: A Research Agenda for the Social Sciences." *Big Data & Society* 9 (1): 20539517221089305.
- Ghasemmaghaei, Maryam, and Nima Kordzadeh. 2024. "Understanding How Algorithmic Injustice Leads to Making Discriminatory Decisions: An Obedience to Authority Perspective." *Information & Management* 61 (2): 103921.
- Green, Ben. 2022. "The Flaws of Policies Requiring Human Oversight of Government Algorithms." *Computer Law & Security Review* 45 (July): 105681. <https://doi.org/10.1016/j.clsr.2022.105681>.
- Greene, Joshua D., Fiery A. Cushman, Lisa E. Stewart, Kelly Lowenberg, Leigh E. Nystrom, and Jonathan D. Cohen. 2016. "Moral Machine: A Platform for Gathering a Human Perspective on Moral Decisions Made by Machine Intelligence." *Science* 352 (6293): 1573–76.

- Grinsztajn, Léo, Edouard Oyallon, and Gaël Varoquaux. 2022. "Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data?" arXiv. <http://arxiv.org/abs/2207.08815>.
- Guerdan, Luke M., Kenneth Holstein, Zhiwei Steven, and Steven Wu. 2022. "Under-Reliance or Misalignment? How Proxy Outcomes Limit Measurement of Appropriate Reliance in AI-Assisted Decision-Making." In . <https://www.semanticscholar.org/paper/Under-reliance-or-misalignment-How-proxy-outcomes-Guerdan-Holstein/317a02a572a450af28352869ffae3d8df0104abd>.
- Hansen, Lauren. 2013. "8 Drivers Who Blindly Followed Their GPS into Disaster." *The Week*. May 8, 2013. <https://theweek.com/articles/464674/8-drivers-who-blindly-followed-gps-into-disaster>.
- Hemmer, Patrick, Max Schemmer, Michael Vössing, and Niklas Kühl. 2021. "Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review." *PACIS 2021 Proceedings*, July. <https://aisel.aisnet.org/pacis2021/78>.
- Hemmer, Paul, Michael Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. 2024. "Complementarity in Human-AI Collaboration: Concept, Sources, and Evidence." *arXiv Preprint arXiv:2404.00029*.
- High-Level Expert Group on Artificial Intelligence. 2020. "Assessment List for Trustworthy Artificial Intelligence (ALTAI)." July 17, 2020. <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-alta-self-assessment>.
- IBM Consulting. 2023. "Artificial Intelligence and a New Era of Human Resources." *IBM Blog* (blog). October 9, 2023. <https://www.ibm.com/blog/artificial-intelligence-and-a-new-era-of-human-resources/>.
- Inkpen, Kori, Sruthi Chappidi, Kaitlyn Mallari, Besmira Nushi, Divya Ramesh, Pietro Michelucci, and others. 2023. "Advancing Human-AI Complementarity: The Impact of User Expertise and Algorithmic Tuning on Joint Decision Making." *ACM Transactions on Computer-Human Interaction* 30 (5): 1–29.
- Jussupow, Ekaterina, Miguel Angel Meza Martínez, Alexander Mädche, and Armin Heinzl. 2021. "Is This System Biased? – How Users React to Gender Bias in an Explainable AI System." In *42nd International Conference on Information Systems*. Association for Information Systems (AIS).
- Kamiran, Faisal, and Toon Calders. 2012. "Data Preprocessing Techniques for Classification without Discrimination." *Knowledge and Information Systems* 33 (1): 1–33. <https://doi.org/10.1007/s10115-011-0463-8>.
- Khan, Arif Ali, Sher Badshah, Peng Liang, Muhammad Waseem, Bilal Khan, Aakash Ahmad, Mahdi Fahmideh, Mahmood Niazi, and Muhammad Azeem Akbar. 2022. "Ethics of AI: A Systematic Literature Review of Principles and Challenges." In *The International Conference on Evaluation and Assessment in Software Engineering 2022*, 383–92. Gothenburg Sweden: ACM. <https://doi.org/10.1145/3530019.3531329>.
- Khan, Tanvir Ahmed. 2023. "Can Unbiased Predictive AI Amplify Bias?" Working Paper 1510. Economics Department, Queen's University. <https://ideas.repec.org/p/qed/wpaper/1510.html>.
- Köbis, Nils, Jean-François Bonnefon, and Iyad Rahwan. 2021. "Bad Machines Corrupt Good Morals." *Nature Human Behaviour* 5 (6): 679–85. <https://doi.org/10.1038/s41562-021-01128-2>.
- Kordzadeh, Nima, and Maryam Ghasemaghaei. 2022. "Algorithmic Bias: Review, Synthesis, and Future Research Directions." *European Journal of Information Systems* 31 (3): 388–409. <https://doi.org/10.1080/0960085X.2021.1927212>.
- Koulu, Riikka. 2020. "Proceduralizing Control and Discretion: Human Oversight in Artificial Intelligence Policy." *Maastricht Journal of European and Comparative Law* 27 (6): 720–35.
- Krawczyk, Michał. 2019. "What Should Be Regarded as Deception in Experimental Economics? Evidence from a Survey of Researchers and Subjects." *Journal of Behavioral and Experimental Economics* 79 (April): 110–18. <https://doi.org/10.1016/j.soec.2019.01.008>.

- Krügel, Sebastian, Andreas Ostermaier, and Matthias Uhl. 2023. "Algorithms as Partners in Crime: A Lesson in Ethics by Design." *Computers in Human Behavior* 138 (January): 107483. <https://doi.org/10.1016/j.chb.2022.107483>.
- Kunda, Z. 1990. "The Case for Motivated Reasoning." *Psychological Bulletin* 108 (3): 480–98. <https://doi.org/10.1037/0033-2909.108.3.480>.
- Lang, Frieder R., Dennis John, Oliver Lüdtke, Jürgen Schupp, and Gert G. Wagner. 2011. "Short Assessment of the Big Five: Robust across Survey Methods except Telephone Interviewing." *Behavior Research Methods* 43 (2): 548–67. <https://doi.org/10.3758/s13428-011-0066-z>.
- Laux, Johann. 2023. "Institutionalised Distrust and Human Oversight of Artificial Intelligence: Towards a Democratic Design of AI Governance under the European Union AI Act." *AI & SOCIETY*, October. <https://doi.org/10.1007/s00146-023-01777-z>.
- Ledford, Heidi. 2019. "Millions of Black People Affected by Racial Bias in Health-Care Algorithms." *Nature* 574 (7780): 608–9. <https://doi.org/10.1038/d41586-019-03228-6>.
- Lee, Julie. 2023. "The Future of AI in Lending." Experian Insights. January 18, 2023. <https://www.experian.com/blogs/insights/future-ai-lending/>.
- Leib, Margarita, Nils C. Köbis, Rainer Michael Rilke, Marloes Hagens, and Bernd Irlenbusch. 2021. "The Corruptive Force of AI-Generated Advice." *arXiv:2102.07536 [Cs, Econ, q-Fin]*, February. <http://arxiv.org/abs/2102.07536>.
- Li, Xuran, Peng Wu, and Jing Su. 2023. "Accurate Fairness: Improving Individual Fairness without Trading Accuracy." *Proceedings of the AAAI Conference on Artificial Intelligence* 37 (12): 14312–20. <https://doi.org/10.1609/aaai.v37i12.26674>.
- Lipton, Zachary C. 2016. "The Mythos of Model Interpretability." *Queue* 16 (3): 31–57.
- Liu, Minghuan, Menghui Zhu, and Weinan Zhang. 2022. "Goal-Conditioned Reinforcement Learning: Problems and Solutions." *arXiv*. <https://doi.org/10.48550/arXiv.2201.08299>.
- Mahmud, Hasan, A.K.M. Najmul Islam, Syed Ishtiaque Ahmed, and Kari Smolander. 2022. "What Influences Algorithmic Decision-Making? A Systematic Literature Review on Algorithm Aversion." *Technological Forecasting and Social Change* 175 (February): 121390. <https://doi.org/10.1016/j.techfore.2021.121390>.
- Mattu, Julia Angwin, Jeff Larson, Lauren Kirchner, Surya. 2016. "Machine Bias." ProPublica. May 23, 2016. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- Maxwell, Winston. 2023. "Meaningful Human Control to Detect Algorithmic Errors." In *Artificial Intelligence Law: Between Sectoral Rules and Comprehensive Regime - Comparative Law Perspectives*, edited by Céline Castets-Renard and Jessica Eynard. Bruylants. <https://hal.science/hal-04026883>.
- McIntosh, Michele J., and Janice M. Morse. 2015. "Situating and Constructing Diversity in Semi-Structured Interviews." *Global Qualitative Nursing Research* 2 (January): 2333393615597674. <https://doi.org/10.1177/2333393615597674>.
- Mehrabi, Nihareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys* 54 (6): 115:1–115:35. <https://doi.org/10.1145/3457607>.
- Mitchell, Shira, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. "Algorithmic Fairness: Choices, Assumptions, and Definitions." *Annual Review of Statistics and Its Application* 8 (1): 141–63. <https://doi.org/10.1146/annurev-statistics-042720-125902>.
- Molinaro, Gaia, and Anne G. E. Collins. 2023. "A Goal-Centric Outlook on Learning." *Trends in Cognitive Sciences* 27 (12): 1150–64. <https://doi.org/10.1016/j.tics.2023.08.011>.
- Morewedge, Carey K. 2022. "Preference for Human, Not Algorithm Aversion." *Trends in Cognitive Sciences*, August, S1364661322001644. <https://doi.org/10.1016/j.tics.2022.07.007>.
- Morozov, Evgeny. 2013. *To Save Everything, Click Here*. <https://www.hachettebookgroup.com/titles/evgeny-morozov/to-save-everything-click-here/9781610393706/?lens=publicaffairs>.

- Orfanoudaki, Agni, Soroush Saghafian, Karen Song, Harini A. Chakker, and Curtiss Cook. 2022. "Algorithm, Human, or the Centaur: How to Enhance Clinical Care?" SSRN Scholarly Paper. Rochester, NY. <https://doi.org/10.2139/ssrn.4302002>.
- Ortmann, Andreas, and Ralph Hertwig. 2002. "The Costs of Deception: Evidence from Psychology." *Experimental Economics* 5 (2): 111–31. <https://doi.org/10.1023/A:1020365204768>.
- Panigutti, Cecilia, Ronan Hamon, Isabelle Hupont, David Fernandez Llorca, Delia Fano Yela, Henrik Junklewitz, Salvatore Scalzo, et al. 2023. "The Role of Explainable AI in the Context of the AI Act." In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1139–50. Chicago IL USA: ACM. <https://doi.org/10.1145/3593013.3594069>.
- Parasuraman, Raja, and Dietrich H. Manzey. 2010. "Complacency and Bias in Human Use of Automation: An Attentional Integration." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 52 (3): 381–410. <https://doi.org/10.1177/0018720810376055>.
- Parasuraman, Raja, Thomas B. Sheridan, and Christopher D. Wickens. 2000. "A Model for Types and Levels of Human Interaction with Automation." *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30 (3): 286–97.
- Parmar, Aakash, Rakesh Katariya, and Vatsal Patel. 2019. "A Review on Random Forest: An Ensemble Classifier." In *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*, edited by Jude Hemanth, Xavier Fernando, Pavel Lafata, and Zubair Baig, 758–63. Cham: Springer International Publishing. [https://doi.org/10.1007/978-3-030-03146-6\\_86](https://doi.org/10.1007/978-3-030-03146-6_86).
- Polit, Denise F, and Cheryl Tatano Beck. 2010. "Generalization in Quantitative and Qualitative Research: Myths and Strategies." *International Journal of Nursing Studies* 47 (11): 1451–58.
- Reinecke, Madeline G., Yiran Mao, Markus Kunesch, Edgar A. Duéñez-Guzmán, Julia Haas, and Joel Z. Leibo. 2023. "The Puzzle of Evaluating Moral Cognition in Artificial Agents." *Cognitive Science* 47 (8): e13315. <https://doi.org/10.1111/cogs.13315>.
- Ribeiro, Marco Tulio Correia. (2016) 2024. "Marcotcr/Lime." JavaScript. <https://github.com/marcotcr/lime>.
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?: Explaining the Predictions of Any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–44. KDD '16. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939778>.
- Roller, Margaret R, and Paul J Lavrakas. 2015. *Applied Qualitative Research Design: A Total Quality Framework Approach*. Guilford Publications.
- Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cédric Langbort. 2014. "Auditing Algorithms : Research Methods for Detecting Discrimination on Internet Platforms." In . <https://www.semanticscholar.org/paper/Auditing-Algorithms-%3A-Research-Methods-for-on-Sandvig-Hamilton/b7227cbd34766655dea10d0437ab10df3a127396>.
- Saunders, Benjamin, Julius Sim, Tom Kingstone, Susan Baker, Jackie Waterfield, Bernadette Bartlam, and others. 2018. "Saturation in Qualitative Research: Exploring Its Conceptualization and Operationalization." *Quality & Quantity* 52: 1893–1907.
- Schemmer, Max, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. "Should I Follow AI-Based Advice? Measuring Appropriate Reliance in Human-AI Decision-Making." arXiv. <http://arxiv.org/abs/2204.06916>.
- Schmuckler, Mark A. 2001. "What Is Ecological Validity? A Dimensional Analysis." *Infancy: The Official Journal of the International Society on Infant Studies* 2 (4): 419–36. [https://doi.org/10.1207/S15327078IN0204\\_02](https://doi.org/10.1207/S15327078IN0204_02).
- Schmude, Tim, Laura Koesten, Timo Möller, and Sebastian Tschiatschek. 2023. "On the Impact of Explanations on Understanding of Algorithmic Decision-Making." In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 959–70.
- Schoeffer, Jakob, Johannes Jakubik, Michael Voessing, Niklas Kuehl, and Gerhard Satzger. 2023. "On the Interdependence of Reliance Behavior and Accuracy in AI-Assisted Decision-Making." arXiv. <https://doi.org/10.48550/arXiv.2304.08804>.

- Schwartz, Reva, Apostol Vassilev, Kristen K. Greene, Lori Perine, Andrew Burt, and Patrick Hall. 2022. "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence." *NIST*, March. <https://www.nist.gov/publications/towards-standard-identifying-and-managing-bias-artificial-intelligence>.
- Scikit-learn. 2024. "Scikit-Learn: Machine Learning in Python." 2024. <https://scikit-learn.org/stable/>.
- Selten, Friso, Marcel Robeert, and Stephan Grimmelikhuijsen. 2023. "Just like I Thought": Street-level Bureaucrats Trust AI Recommendations If They Confirm Their Professional Judgment." *Public Administration Review* 83 (2): 263–78. <https://doi.org/10.1111/puar.13602>.
- Shneiderman, Ben. 2020. "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy." *International Journal of Human-Computer Interaction* 36 (6): 495–504.
- Slavkovik, Marija. 2023. "Mythical Ethical Principles for AI and How to Attain Them." In *Human-Centered Artificial Intelligence*, edited by Mohamed Chetouani, Virginia Dignum, Paul Lukowicz, and Carles Sierra, 13500:275–303. Lecture Notes in Computer Science. Cham: Springer International Publishing. [https://doi.org/10.1007/978-3-031-24349-3\\_15](https://doi.org/10.1007/978-3-031-24349-3_15).
- Smith, Jonathan A. 1995. "Semi Structured Interviewing and Qualitative Analysis." In *Rethinking Methods in Psychology*, edited by Jonathan A. Smith, R. Harre, and L. Van Langenhove, 9–26. Sage Publications. <https://uk.sagepub.com/en-gb/eur/rethinking-methods-in-psychology/book204294>.
- Srivastava, Prachi, and Nick Hopwood. 2009. "A Practical Iterative Framework for Qualitative Data Analysis." *International Journal of Qualitative Methods* 8 (1): 76–84.
- Strauss, Anselm L. 1987. *Qualitative Analysis for Social Scientists*. Qualitative Analysis for Social Scientists. New York, NY, US: Cambridge University Press. <https://doi.org/10.1017/CBO9780511557842>.
- Tolan, Songül. 2019. "Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges." *arXiv:1901.04730 [Cs, Stat]*, January. <http://arxiv.org/abs/1901.04730>.
- Tsamados, Antonios, Nikita Aggarwal, Josh Cowls, Jessica Morley, Huw Roberts, Mariarosaria Taddeo, and Luciano Floridi. 2021. *The Ethics of Algorithms: Key Problems and Solutions*. Springer.
- Tsing, Anna Lowenhaupt. 2004. *Friction: An Ethnography of Global Connection*. Illustrated édition. Princeton, NJ: Princeton University Press.
- Tversky, Amos, and Daniel Kahneman. 1986. "Rational Choice and the Framing of Decisions." *The Journal of Business* 59 (4): S251–78.
- UNESCO. 2021. "Recommendation on the Ethics of Artificial Intelligence." <https://unesdoc.unesco.org/ark:/48223/pf0000381137>.
- Vaismoradi, Mojtaba, Jordan Jones, Hannele Turunen, and Sherrill Snelgrove. 2016. "Theme Development in Qualitative Content Analysis and Thematic Analysis." *Journal of Nursing Education and Practice* 6 (5).
- Vasileiou, Konstantina, Julie Barnett, Susan Thorpe, and Terry Young. 2018. "Characterising and Justifying Sample Size Sufficiency in Interview-Based Studies: Systematic Analysis of Qualitative Health Research over a 15-Year Period." *BMC Medical Research Methodology* 18 (1): 148. <https://doi.org/10.1186/s12874-018-0594-7>.
- Vaughn, Lisa M., and Farrah Jacquez. 2020. "Participatory Research Methods – Choice Points in the Research Process." *Journal of Participatory Research Methods* 1 (1). <https://doi.org/10.35844/001c.13244>.
- Veksler, Vladislav D., Wayne D. Gray, and Michael J. Schoelles. 2013. "Goal-Proximity Decision-Making." *Cognitive Science* 37 (4): 757–74. <https://doi.org/10.1111/cogs.12034>.
- Vlasceanu, Madalina, and David M Armodio. 2022. "Propagation of Societal Gender Inequality by Internet Search Algorithms." *Proceedings of the National Academy of Sciences* 119 (29): e2204529119.
- Wang, Clarice, Kathryn Wang, Andrew Y. Bian, Rashidul Islam, Kamrun Naher Keya, James Foulds, and Shimei Pan. 2023. "When Biased Humans Meet Debiased AI: A Case Study in College Major Recommendation." *ACM Transactions on Interactive Intelligent Systems* 13 (3): 1–28. <https://doi.org/10.1145/3611313>.

- Zick, Andreas, Beate Küpper, and Andreas Hövermann. 2011. *Intolerance, Prejudice and Discrimination. A European Report*. Universität Tübingen.
- Zliobaite, Indre. 2017. "Measuring Discrimination in Algorithmic Decision Making." <https://core.ac.uk/reader/245131330>.

## List of abbreviations and definitions

Abbreviations	Definitions
AI	Artificial Intelligence
ALTAI	Assessment List for Trustworthy Artificial Intelligence
DE	Germany
DSS	Decision Support System
EC	European Commission
EWP	Expert workshop participant
FIN	Finance
HITL	Human-in-the-Loop
HR	Human Resources
IT	Italy

## List of figures

Figure 1 Exploring the combination of human and AI biases.....	9
Figure 2 Human and AI decisional background and their interplay.....	10
Figure 3 Schematic process of evaluation and decision, with and without AI.....	15
Figure 4 Pipelines followed to obtain AI-based Decision Support System's outputs plus explanations for deciders. Top pipeline corresponds to the generation of "generic" outputs and explanations; Bottom pipeline represents the process to obtain "fair" outputs .....	19
Figure 5 Chronology of the deciders' experiment.....	21
Figure 6 Interface used for the examination of algorithmic biased during the interviews, German version.....	28
Figure 7 Screenshot from the videocall, sharing the screen with the Miro board activity for the workshop with participants.....	29
Figure 8 Workshop with experts during the presentation of the outcome of activity 1 .....	31
Figure 9 Workshop with experts, output of the tools for activity 2 for the mortgage scenario.....	32
Figure 10 Deciders' preferences among applicants, by sector.....	38
Table 16 Deciders' performance vs. AI performance.....	39
Figure 11 Impact of applicant characteristics on selection, treatment without AI .....	41
Figure 12 Likelihood to choose a candidate as a function of AI recommendation for or against that candidate.....	42
Figure 13 Influence of AI grades on choice.....	43
Figure 14 Gender and country discrimination, by AI type.....	44
Figure 15 Effect of AIs on gender discrimination by gender of decider .....	45

Figure 16: Effect of AIs on country discrimination by country of the decider .....	46
Figure 17 Bias as a function of own discriminatory preferences, for gender and nationality .....	47
Figure 18 Bias as a function of own discriminatory preferences, for all applicant characteristics....	48
Figure 19 ALTAI Framework .....	57
Figure 20 Annotation scheme for the analysis of the interviews with the professionals.....	58
Figure 21 Screenshot of the Miro Activity run during the workshop, in specific regarding the discussion on characteristics and ecological validity of the experiment.....	62
Figure 22 Workshop with experts during the group work discussion.....	63
Figure 23 One of the tarot cards that represent the “future” definition of fairness in the AI+Human decision-making process.....	66
Figure 24 Workshop with policy-makers in the EU Policy Lab, discussing opportunities and needs in regards to Human Oversight .....	69
Figure 22 Preferences of the AI, by sector and type of AI.....	100
Figure 23 Deciders' preferences among applicants, by gender and country.....	101
Figure 24 Correlation between decider characteristics, preferences, and prejudices.....	102

## List of tables

Table 1 Accuracy metrics for each DSS scenario .....	19
Table 2 Deciders' sample distribution by sector, country, and treatment.....	21
Table 2 Presentation of decider's own preferences, example.....	22
Table 3 The decision interface showing explanations for the AI recommendation, example.....	23
Table 4 Phase 1 of the interview: Questions that contribute to the identification of contextual information.....	27
Table 5 Phase 2 of the interview: Questions to support the participants' reflection about their interaction with the algorithm.....	27
Table 6 Phase 3 of interview: Questions to support participant reflection about the explanations provided during the activity in phase 2.....	28
Table 7 Structure of the 1-day workshop with Experts .....	31
Table 8 Socio-demographics, deciders .....	33
Table 9 Work environment and experience.....	34
Table 10 Experience with data and DSS.....	35
Table 11 Company diversity .....	35
Table 12 Mode of decisions.....	36
Table 13 Evaluation of the AI .....	37
Table 14 Prejudice and views on discrimination.....	39
Table 15 Deciders' performance vs. AI performance .....	39
Table 16 Perceptions of the AI by deciders .....	49
Table 17 Distribution of interview participants .....	50
Table 18 Demographics of interview participants, with participants codes.....	50
Table 19 Distribution of workshop participants .....	51
Table 20 List of Experts with their specialization/discipline.....	52
Table 21 Iterative procedure of data analysis in three steps .....	53
Table 22 Main findings and opportunities to improve human oversight.....	70
Table 23 Factors in the choice among applicants, treatment without AI .....	102

## Annexes

### Annex 1. Sample characteristics vs. quotas

#### Recipients

	Target		Completed	
	Germany	Italy	Germany	Italy
<b>Age</b>				
18-24	12%	11%	12%	11%
25-34	20%	17%	17%	18%
35-44	20%	20%	19%	21%
45-54	22%	26%	21%	22%
55-65	26%	25%	32%	28%
<b>Gender</b>				
Male	51%	50%	46%	49%
Female	49%	50%	54%	51%
<b>Regions Germany</b>				
Baden-Württemberg	14%	-	12%	-
Bayern	16%	-	14%	-
Berlin	5%	-	5%	-
Brandenburg	3%	-	2%	-
Bremen	1%	-	1%	-
Hamburg	2%	-	2%	-
Hessen	8%	-	8%	-
Mecklenburg-Vorpommern	2%	-	1%	-
Niedersachsen	10%	-	11%	-
Nordrhein-Westfalen	22%	-	24%	-
Rheinland-Pfalz	5%	-	6%	-
Saarland	1%	-	1%	-
Sachsen	5%	-	4%	-
Sachsen-Anhalt	3%	-	3%	-
Schleswig-Holstein	3%	-	4%	-
Thüringen	2%	-	2%	-

Regions Italy				
	-	7%	-	7%
Piemonte	-	7%	-	7%
Valle d'Aosta/Vallée d'Aoste	-	0,2%	-	0,4%
Liguria	-	2%	-	2%
Lombardia	-	17%	-	17%
Provincia Autonoma di Bolzano/Bozen	-	1%	-	0%
Provincia Autonoma di Trento	-	1%	-	0%
Veneto	-	8%	-	8%
Friuli-Venezia Giulia	-	2%	-	2%
Emilia-Romagna	-	7%	-	7%
Toscana	-	6%	-	7%
Umbria	-	1%	-	1%
Marche	-	3%	-	3%
Lazio	-	10%	-	11%
Abruzzo	-	2%	-	2%
Molise	-	1%	-	0,4%
Campania	-	10%	-	10%
Puglia	-	7%	-	7%
Basilicata	-	1%	-	0,4%
Calabria	-	3%	-	3%
Sicilia	-	8%	-	8%
Sardegna		3%	-	3%

## Deciders

Age	HR		Retail Banking	
	Germany	Italy	Germany	Italy
18-24	2%	5%	-	3%
25-34	17%	26%	28%	20%
35-44	63%	48%	54%	54%
45-54	13%	16%	17%	17%

55-65	5%	5%	2%	5%
Gender				
Male	48%	50%	52%	48%
Female	52%	50%	48%	52%
Regions Germany				
Baden-Württemberg	10%	-	11%	-
Bayern	19%	-	14%	-
Berlin	18%	-	12%	-
Brandenburg	3%	-	3%	-
Bremen	2%	-	3%	-
Hamburg	4%	-	6%	-
Hessen	9%	-	14%	-
Mecklenburg-Vorpommern	1%	-	1%	-
Niedersachsen	5%	-	4%	-
Nordrhein-Westfalen	14%	-	20%	-
Rheinland-Pfalz	5%	-	3%	-
Saarland	1%	-	2%	-
Sachsen-Anhalt	2%	-	4%	-
Sachsen-Anhalt	1%	-	0.30%	-
Schleswig-Holstein	5%	-	2%	-
Thüringen	1%	-	1%	-
Regions Italy				
Piemonte	-	7%	-	6%
Valle d'Aosta/Vallée d'Aoste	-	1%	-	0.30%
Liguria	-	4%	-	2%
Lombardia	-	20%	-	21%
Provincia Autonoma di Bolzano/Bozen	-	1%	-	2%
Provincia Autonoma di Trento	-	3%	-	1%
Veneto	-	4%	-	3%
Friuli-Venezia Giulia	-	2%	-	3%
Emilia-Romagna	-	7%	-	7%
Toscana	-	6%	-	2%
Umbria	-	3%	-	4%

Marche	-	3%	-	3%
Lazio	-	10%	-	16%
Abruzzo	-	2%	-	3%
Molise	-	1%	-	2%
Campania	-	14%	-	9%
Puglia	-	4%	-	6%
Basilicata	-	2%	-	0.30%
Calabria	-	1%	-	2%
Sicilia	-	5%	-	6%
Sardegna	-	3%	-	1%

## Annex 2. Variables collected

### Recipients

Variable Name	Labels	Values
Respondent_Serial	Unique identifier	
Country	Country	1 Germany 2 Italy
Device	The device used in the latest access of the survey link	1 Laptop/PC 2 Smartphone 3 Tablet 4 SmartTV 5 None of the above
D1	Gender	1 Male 2 Female
GENDER_NonBinary	Are you...?	1 Male 2 Female 3 Another gender 4 Prefer not to say
resp_age	RespondentAge	
AgeCat	Age in Categories	1 18-24 2 25-34 3 35-44 4 45-54 5 55-65
D3_DE	In which region do you live?	1 Baden-Württemberg 2 Bayern 3 Berlin 4 Brandenburg 5 Bremen 6 Hamburg 7 Hessen 8 Mecklenburg-Vorpommern 9 Niedersachsen 10 Nordrhein-Westfalen 11 Rheinland-Pfalz 12 Saarland 13 Sachsen 14 Sachsen-Anhalt 15 Schleswig-Holstein 16 Thüringen 998 Don't know 999 Prefer not to answer
D3_IT	In which region do you live?	1 Piemonte 2 Valle d'Aosta/Vallée d'Aoste 3 Liguria 4 Lombardia 5 Provincia Autonoma di Bolzano/Bozen 6 Provincia Autonoma di Trento 7 Veneto 8 Friuli-Venezia Giulia 9 Emilia-Romagna 10 Toscana 11 Umbria 12 Marche 13 Lazio 14 Abruzzo 15 Molise 16 Campania 17 Puglia 18 Basilicata 19 Calabria 20 Sicilia 21 Sardegna 998 Don't know 999 Prefer not to answer
QTEST1	What is the sum of below numbers? 6+1+6+7 =	

Intro9	How many of those sums do you think you answered correctly?	
Completed_Sums	Total answered sums	
Correct_Sums	Total correctly answered sums	
Intro12_Input	What do you have to do?	
Intro12_Retry	If you repay [insert R] points, then you get [insert 300-R] points and the banker gets [insert R] points.	1 Yes 2 No
QCHECK2_Input	Let us now check that you understood the instructions for this second part correctly. Please choose a number between 0 and 300	
QCHECK2_1	How many points do you get?	
QCHECK2_2	How much does the banker get?	
QCHECK2_3	How much does the banker get if he or she decides not to lend you?	1 0 points 2 100 points 3 200 points 4 300 points
QTEST3	How many points do you choose to give back to the banker?	
Q1	What do you think this experiment was about?	
Q1_translated	What do you think this experiment was about?	
Q2_1	To what extent do you agree with the below statements about the first part of the experiment where you summed numbers? - I tried to do my best	1 Totally disagree 2 Tend to disagree 3 Tend to agree 4 Totally agree
Q2_2	To what extent do you agree with the below statements about the first part of the experiment where you summed numbers? - I enjoyed the task	1 Totally disagree 2 Tend to disagree 3 Tend to agree 4 Totally agree
Q2_3	To what extent do you agree with the below statements about the first part of the experiment where you summed numbers? - The task was difficult	1 Totally disagree 2 Tend to disagree 3 Tend to agree 4 Totally agree
Q2_4	To what extent do you agree with the below statements about the first part of the experiment where you summed numbers? - It was hard for me to understand what I had to do	1 Totally disagree 2 Tend to disagree 3 Tend to agree 4 Totally agree
Q2_5	To what extent do you agree with the below statements about the first part of the experiment where you summed numbers? - I think I did well in this task	1 Totally disagree 2 Tend to disagree 3 Tend to agree 4 Totally agree
Q3	Please choose the option that best describes your decision about the second part of the experiment where you had to repay a loan.	1 I kept as many points for me as possible 2 I simply paid back the loan but not more 3 I tried to share the earnings from the loan equally 4 I wanted to reward the banker for lending to me See options at the end of this table
D4	What is the highest level of school you have completed, or the highest degree you have received?	
ISCED	What is the highest level of school you have completed, or the highest degree you have received?	1 Low 2 Middle 3 High

ISCED2	What is the highest level of school you have completed, or the highest degree you have received?	1 Low 2 Middle 3 High
D5	Do you have the German / Italian nationality?	1 Yes 2 No
D6	What is your current occupation?	1 Student 2 Working full time 3 Working part time 4 Unemployed 5 Retired 6 Looking for a job 7 Housewife/houseman 98 Other
D7	What sector are you employed in?	1 Public 2 Private 3 Self-employed
D8	How large is the company you work in?	1 Less than 10 employees 2 10-49 employees 3 50-249 employees 4 More than 250 employees 99 Don't know
D9_DE	Could you please indicate your household's monthly income (that is, after income taxes have been paid)?	1 Less than 1299 euro 2 between 1300 and 1749 euro 3 between 1750 and 2199 euro 4 between 2200 and 2849 euro 5 2850 euro or more 98 Don't know 99 Prefer not to answer
D9_IT	Could you please indicate your household's monthly income (that is, after income taxes have been paid)?	1 Less than 849 euro 2 between 850 and 1249 euro 3 between 1250 and 1649 euro 4 between 1650 and 2199 euro 5 2200 euro or more 98 Don't know 99 Prefer not to answer
D10	What social class do you feel you belong to?	1 Working class 2 Middle class 3 Upper class 99 Prefer not to say
D11_1	Do you generally...	1 1 - Try to avoid taking risks 2 2 3 3 4 4 - Are comfortable with taking risks
D11_2	Do you think...	1 1 - Incomes should be made more equal 2 2 3 3 4 4 - Incomes should depend more on individual effort
D11_3	Do you think...	1 1 - Competition is good, it brings the best out of people 2 2 3 3 4 4 - Competition is bad, it brings the worst out of people
D11_4	Do you think...	1 1 - What happens to you is your own doing 2 2 3 3 4 4 - You have little influence over what happens to you
D11_5	Would you say that...	1 1 - Most people can be trusted 2 2 3 3 4 4 - You can't be too careful in dealing with people
D12_1	To what extent do you agree with the below statements? - Worries a lot	1 Totally disagree 2 Tend to disagree 3 Tend to agree 4 Totally agree
D12_2	To what extent do you agree with the below statements? - Gets nervous easily	1 Totally disagree 2 Tend to disagree 3 Tend to agree 4 Totally agree
D12_3	To what extent do you agree with the below statements? - Remains calm in tense situations	1 Totally disagree 2 Tend to disagree 3 Tend to agree 4 Totally agree
D12_4	To what extent do you agree with the below statements? - Is talkative	1 Totally disagree 2 Tend to disagree 3 Tend to agree 4 Totally agree
D12_5	To what extent do you agree with the below statements? - Is outgoing, sociable	1 Totally disagree 2 Tend to disagree 3 Tend to agree 4 Totally agree
D12_6	To what extent do you agree with the below statements? - Is reserved	1 Totally disagree 2 Tend to disagree 3 Tend to agree 4 Totally agree

D12_7	To what extent do you agree with the below statements? - Is original, comes up with new ideas	1 Totally disagree 2 Tend to disagree 3 Tend to agree 4 Totally agree
D12_8	To what extent do you agree with the below statements? - Values artistic, aesthetic experiences	1 Totally disagree 2 Tend to disagree 3 Tend to agree 4 Totally agree
D12_9	To what extent do you agree with the below statements? - Has an active imagination	1 Totally disagree 2 Tend to disagree 3 Tend to agree 4 Totally agree
D12_10	To what extent do you agree with the below statements? - Is sometimes rude to others	1 Totally disagree 2 Tend to disagree 3 Tend to agree 4 Totally agree
D12_11	To what extent do you agree with the below statements? - Has a forgiving nature	1 Totally disagree 2 Tend to disagree 3 Tend to agree 4 Totally agree
D12_12	To what extent do you agree with the below statements? - Is considerate and kind to almost everyone	1 Totally disagree 2 Tend to disagree 3 Tend to agree 4 Totally agree
D12_13	To what extent do you agree with the below statements? - Does a thorough job	1 Totally disagree 2 Tend to disagree 3 Tend to agree 4 Totally agree
D12_14	To what extent do you agree with the below statements? - Tends to be lazy	1 Totally disagree 2 Tend to disagree 3 Tend to agree 4 Totally agree
D12_15	To what extent do you agree with the below statements? - Does things efficiently	1 Totally disagree 2 Tend to disagree 3 Tend to agree 4 Totally agree

Options in response to D4 are:

- in Germany: 1 Kein Schulabschluss, ich habe die Schule vor Erreichen des 15. Lebensjahres verlassen  
2 Hauptschul-/POS; ohne beruflichen Abschluss 3 Hauptschul-/POS; Anlernausbildung, Berufliches Praktikum 4 Hauptschul-/POS; Berufsvorbereitungsjahr 5 Ohne Schulabschluss; Anlernausbildung; Berufliches Praktikum 6 Ohne Schulabschluss; Berufsvorbereitungsjahr 7 Fachhochschulreife/Hochschulreife; ohne beruflichen Abschluss 8 Realschulabschluss mit Abschluss einer Lehrausbildung/Polytechnische Oberschule Abschluss nach 10 Jahren 9 Berufsqualifizierender Abschluss an Berufsfachschulen/Kollegschen 10 Abschluss einer 1-jährigen Schule des Gesundheitswesens 11 Abschluss des Vorbereitungsdienstes für den mittleren Dienst in der öffentlichen Verwaltung 12 Fachhochschulreife/Hochschulreife und Abschluss einer Lehrausbildung 13 Fachhochschulreife/Hochschulreife und Berufsqualifizierender Abschluss an 14 Berufsfachschulen/Kollegschen, Abschluss einer einjährigen Schule des Gesundheitswesens 15 Fachhochschulreife/Hochschulreife und Abschluss des Vorbereitungsdienstes für den mittleren Dienst in der öffentlichen V 16 Fachhochschulabschluss (auch Ingenieurschulabschluss, Diplom (FH), Bachelor-/Masterabschluss an Fachhochschulen, ohne Ab 17 Hochschulabschluss (Diplom (Universität) und entsprechende Abschlussprüfungen,Künstlerischer Abschluss, Bachelor-/ Maste 18 Meister-/Technikerausbildung oder gleichwertiger Fachschulabschluss, Abschluss einer 2- oder 3- jährigen Schule des Gesu 19 Abschluss einer Verwaltungsfachhochschule (Diplom, Bachelor, Master an Verwaltungsfachhochschulen) 20 Abschluss der Fachschule der ehemaligen DDR 21 Promotion
- In Italy: 22 Nessuna scuola 23 Elementare non conclusa 24 Elementare con licenza 25 Media inferiore non conclusa 26 Media inferiore con licenza 27 Media superiore non conclusa 28 Media superiore con diploma 29 Diploma universitario / extra-universitario 30 Università ma senza laurea 31 Laurea di primo livello/laurea triennale 32 Laurea di secondo livello/laurea specialistica 33 Laurea specialistica a ciclo unico 34 Diploma di laurea (vecchio ordinamento)
- In both countries, there are also options: 98 Don't know 99 Prefer not to answer

The Big 5 personality questionnaire is summarized into 5 dimensions as follows:

- neuroticism=(D12\_1+D12\_2+5-D12\_3)/3
- extroversion=(D12\_4+D12\_5+5-D12\_6)/3
- openness=(D12\_7+D12\_8+D12\_9)/3
- agreeability=(5-D12\_10+D12\_11+D12\_12)/3
- conscientiousness=(D12\_13+5-D12\_14+D12\_15)/3

## Deciders

We show the wordings for the HR experiment. The wordings in the Banking experiment were the same, but replacing "hire" with "lend", "job" with "loan" and "HR" with "Banking".

Variable Name	Labels	Values
Respondent_Serial	Unique identifier	
Country	Country	1 Germany 2 Italy
Device	The device used in the latest access of the survey link	1 Laptop/PC 2 Smartphone 3 Tablet 4 SmartTV 5 None of the above
GROUP	GROUP	1 Employer 2 Lender
Treatment	Treatment	1 Treatment 1: deciders make decisions on their own 2 Treatment 2: deciders get recommendations from an "efficient" AI 3 Treatment 3: deciders get recommendations from a "fair" AI
Pool	Pool	From 1 to 22
Order1	Order 12-1	0 No 1 Yes
Order2	Order 1-12	0 No 1 Yes
SCR1	In which of the following sectors are you currently working?	1 Human Resources Management 2 Retail Banking 3 IT Services 4 Communications 5 Wholesale and retail trade 6 Manufacturing 7 Construction 8 Transportation 9 Food services 10 Other
resp_age	RespondentAge	
QUOTAGERANGE	RespondentAge	1 18-24 2 25-34 3 35-44 4 45-54 5 55-65
resp_gender	Are you...?	1 Male 2 Female
GENDER_NonBinary	Are you...?	1 Male 2 Female 3 In another way 4 Prefer not to answer
D3_DE	In which region do you live?	1 Baden-Württemberg 2 Bayern 3 Berlin 4 Brandenburg 5 Bremen 6 Hamburg 7 Hessen 8 Mecklenburg-Vorpommern 9 Niedersachsen 10 Nordrhein-Westfalen 11

		Rheinland-Pfalz 12 Saarland 13 Sachsen 14 Sachsen-Anhalt 15 Schleswig-Holstein 16 Thüringen 998 I don't know 999 Prefer not to answer
D3_IT	In which region do you live?	1 Piemonte 2 Valle d'Aosta/Vallée d'Aoste 3 Liguria 4 Lombardia 5 Provincia Autonoma di Bolzano/Bozen 6 Provincia Autonoma di Trento 7 Veneto 8 Friuli-Venezia Giulia 9 Emilia-Romagna 10 Toscana 11 Umbria 12 Marche 13 Lazio 14 Abruzzo 15 Molise 16 Campania 17 Puglia 18 Basilicata 19 Calabria 20 Sicilia 21 Sardegna 998 I don't know 999 Prefer not to answer
Check1_1	What kind of task do employees have to do?	1 Describe images 2 Add up numbers 3 Transcribe a text
Check1_2	How much do you get for each correct sum computed by your employee?	1 4 points 2 5 points 3 100 points
Check1_3	What is the wage you pay the person you hire?	1 4 points 2 100 points 3 The person you hire receives no wage
Check1_4	How much does a job applicant earn if you do not select him or her as an employee?	1 100 points 2 4 points 3 0 points
QTest1	How to make a choice?	1 I want to hire A 2 I want to hire B
QRANK1_1	Please tell us which personal characteristics you think are most important when selecting among job applicants. - Gender	1 High Importance 2 Moderate importance 3 Low importance 4 Irrelevant
QRANK1_2	Please tell us which personal characteristics you think are most important when selecting among job applicants. - Age	1 High Importance 2 Moderate importance 3 Low importance 4 Irrelevant
QRANK1_3	Please tell us which personal characteristics you think are most important when selecting among job applicants. - Nationality	1 High Importance 2 Moderate importance 3 Low importance 4 Irrelevant

QRANK1_4	Please tell us which personal characteristics you think are most important when selecting among job applicants. - Level of education	1 High Importance 2 Moderate importance 3 Low importance 4 Irrelevant
QRANK1_5	Please tell us which personal characteristics you think are most important when selecting among job applicants. - Income	1 High Importance 2 Moderate importance 3 Low importance 4 Irrelevant
QRANK1_6	Please tell us which personal characteristics you think are most important when selecting among job applicants. - Interview	1 High Importance 2 Moderate importance 3 Low importance 4 Irrelevant
QRank2_Gender	Please now tell us what type of job applicant you would favour when deciding who to hire?	1 Women 2 Men
QRank2_Age	Please now tell us what type of job applicant you would favour when deciding who to hire?	1 18-34 year old 2 35-54 year old 3 55-65 year old
QRank2_Country	Please now tell us what type of job applicant you would favour when deciding who to hire?	1 Germany 2 Italy
QRank2_Education	Please now tell us what type of job applicant you would favour when deciding who to hire?	1 Low 2 Medium 3 High
QRank2_Income	Please now tell us what type of job applicant you would favour when deciding who to hire?	1 Low 2 Medium 3 High 4 Unknown
QRank2_Interview	Please now tell us what type of job applicant you would favour when deciding who to hire?	1 Bad 2 OK 3 Good 4 Very Good
Intro13_1	At this point, we would like to know if you have difficulties seeing the gradation of colours from	1 Yes, I have difficulty seeing the gradation of colours above 2 No, I have no problem seeing the gradation of colours above

	dark green to dark red above.	
QCheck2_1	How was the DSS programmed?	1 Based on data about past job applicants and their performance in the summing task 2 Based on hiring decisions by other HR managers
QCheck2_2	Does the DSS discriminate across job applicants based on protected characteristics such as gender or nationality?	1 Yes 2 No
QCheck2_3	Do you have to choose based on the grade given by the DSS?	1 Yes 2 No
Q1	What do you think this survey was about?	
Q1_translated	What do you think this survey was about?	
Q2	Please explain how you could have made better decisions, and what prevented you from doing so.	
Q2_translated	Please explain how you could have made better decisions, and what prevented you from doing so.	
Q3	What was more important for you?	1 Choose job applicants that are the most likely to perform well 2 2 3 3 4 Making sure that everyone has a fair chance to get selected
Q4	In what way did you try to make decisions?	1 I tried to make rational decisions 2 2 3 3 4 I trusted my instinct more
Q5	What was your priority when making choices?	1 Making fast decisions 2 2 3 3 4 Making correct decisions
Q6	Overall, how confident were you that you made the right choice?	1 Very confident 2 Fairly confident 3 Not very confident 4 Not at all confident 99 I don't know
Q7_1	How would you rate the general attitude and behavior of men and women? - Honesty	1 Men are generally more honest than women 2 Men and women are generally equally honest 3 Women are generally more honest than men

Q7_2	How would you rate the general attitude and behavior of men and women? - Hard-work	1 Men are generally more hard-working than women 2 Men and women are generally equally hard-working 3 Women are generally more hard-working than men
Q7_3	How would you rate the general attitude and behavior of men and women? - Reliability	1 Men are generally more reliable than women 2 Men and women are generally equally reliable 3 Women are generally more reliable than men
Q7_4	How would you rate the general attitude and behavior of men and women? - Performance	1 Men generally perform better than women 2 Men and women generally perform equally well 3 Women generally perform better than men
Q8_1	How would you rate the general attitude and behavior of Germans and Italians? - Honesty	1 Germans are generally more honest than Italians 2 Germans and Italians are generally equally honest 3 Italians are generally more honest than Germans
Q8_2	How would you rate the general attitude and behavior of Germans and Italians? - Hard-work	1 Germans are generally more hard-working than Italians 2 Germans and Italians are generally equally hard-working 3 Italians are generally more hard-working than Germans
Q8_3	How would you rate the general attitude and behavior of Germans and Italians? - Reliability	1 Germans are generally more reliable than Italians 2 Germans and Italians are generally equally reliable 3 Italians are generally more reliable than Germans
Q8_4	How would you rate the general attitude and behavior of Germans and Italians? - Performance	1 Germans generally perform better than Italians 2 Germans and Italians generally perform equally well 3 Italians generally perform better than Germans
Q9	In this survey, do you think it was OK to choose a job applicant based on their gender?	1 No, never 2 No, rarely 3 Yes, sometimes 4 Yes, always
Q10	In this survey, do you think it was OK to choose a job applicant based on their nationality?	1 No, never 2 No, rarely 3 Yes, sometimes 4 Yes, always
Q11	Did you rely on the DSS when making choices?	1 Yes, to a large extent 2 Yes, somewhat 3 No, not really 4 No, not at all
Q12	Did you understand how the DSS graded job applicants?	1 Yes, to a large extent 2 Yes, somewhat 3 No, not really 4 No, not at all
Q13	Was the DSS fair when grading job applicants?	1 Yes, to a large extent 2 Yes, somewhat 3 No, not really 4 No, not at all

Q14	Was the DSS accurate when grading job applicants?	1 Yes, to a large extent 2 Yes, somewhat 3 No, not really 4 No, not at all
Q15	How long have you been working in Human Resources Management?	1 Less than one year 2 Between one and two years 3 Between three and five years 4 More than five years
Q16	In your current position how many employees report to you?	1 In my current position no one reports to me 2 Between 1 and 5 employees 3 Between 6 and 10 employees 4 Between 11 and 20 employees 5 More than 20 employees 98 I don't know
Q17	How large is the company you work in?	1 Less than 10 employees 2 10-49 employees 3 50-249 employees 4 More than 250 employees 98 I don't know
Q18	How often are you dealing with data and statistics in your job?	1 Very often 2 Sometimes 3 Rarely 4 Never
Q19	How often do you use DSS when hiring / lending at your organisation?	1 Very often 2 Often 3 Rarely 4 Never 98 I don't know
Q20	Can you give more details on the type of DSS you use?	
Q20_translated	Can you give more details on the type of DSS you use?	
Q21	Is there diversity in terms of gender, age and ethnicity, disability status, etc... in the workforce at your company?	1 Yes, there is a lot of diversity 2 Yes, there is some diversity 3 No, there is not much diversity 4 No, there is no diversity
Q22	Are there policies in place to ensure diversity in the workforce at your company?	1 Yes 2 No 98 I don't know
Q23	How well does your organization implement its diversity policies?	1 Very well 2 Well 3 Average 4 Badly 5 Very badly 98 I don't know
D4	What is the highest level of school you have completed, or the highest degree you have received?	See text at the bottom of the table
ISCED	What is the highest level of school you have	1 Low 2 Middle 3 High

	completed, or the highest degree you have received?	
ISCED2	What is the highest level of school you have completed, or the highest degree you have received?	1 Low 2 Middle 3 High
D5	Do you have the German / Italian nationality?	1 Yes 2 No
D6_DE	Could you please indicate your household's monthly income (that is, after income taxes have been paid)?	1 Less than 1299 euro 2 between 1300 and 1749 euro 3 between 1750 and 2199 euro 4 between 2200 and 2849 euro 5 2850 euro or more 98 I don't know 99 Prefer not to answer
D6_IT	Could you please indicate your household's monthly income (that is, after income taxes have been paid)?	1 Less than 849 euro 2 between 850 and 1249 euro 3 between 1250 and 1649 euro 4 between 1650 and 2199 euro 5 2200 euro or more 98 I don't know 99 Prefer not to answer
D7	What social class do you feel you belong to?	1 Working class 2 Middle class 3 Upper class 99 Prefer not to say
QCONSENT_clickedCounter	Clicked counter for QConsent	0 No 1 Yes
Extra_Incentive	Additional amount of euros the respondent will receive (above points * 0.043)	

### Choice variables, deciders

Variable Name	Labels	Values
QT	Please choose your preferred job applicant.	1 Applicant1 2 Applicant2
QT_group	Group Excel input	
gender_1	Gender applicant 1	1 Women 2 Men
gender_2	Gender applicant 2	1 Women 2 Men
age_1	Age applicant 1	1 18-34 year old 2 35-54 year old 3 55-65 year old
age_2	Age applicant 2	1 18-34 year old 2 35-54 year old 3 55-65 year old
education_1	Education applicant 1	1 Low 2 Medium 3 High
education_2	Education applicant 2	1 Low 2 Medium 3 High

country_1	Country applicant 1	1 Germany 2 Italy
country_2	Country applicant 2	1 Germany 2 Italy
income_1	Income applicant 1	1 Low 2 Medium 3 High 4 Unknown
income_2	Income applicant 2	1 Low 2 Medium 3 High 4 Unknown
interview_1	Interview applicant 1	1 Bad 2 OK 3 Good 4 Very Good
interview_2	Interview applicant 2	1 Bad 2 OK 3 Good 4 Very Good
gender_grade_1	Gender grade applicant 1	From 1 to 5, shown as --,-,=,+,+
gender_grade_2	Gender grade applicant 2	idem
age_grade_1	Age grade applicant 1	idem
age_grade_2	Age grade applicant 2	idem
education_grade_1	Education grade applicant 1	idem
education_grade_2	Education grade applicant 2	idem
country_grade_1	Country grade applicant 1	idem
country_grade_2	Country grade applicant 2	idem
income_grade_1	Income grade applicant 1	idem
income_grade_2	Income grade applicant 2	idem
interview_grade_1	Interview grade applicant 1	idem
interview_grade_2	Interview grade applicant 2	idem
overall_grade_1	Overall grade applicant 1	idem
overall_grade_2	Overall grade applicant 2	idem

### Annex 3. Preferences of the AI-based DSS

Preferences of the AI DSS were shown as in the following tables, by sector and type of AI.

#### 1) HR discriminatory

Variables	Importance	Preferred type
Gender	High	Male
Age	Middle	35-54
Nationality	Low	German
Income	Middle	High
Education	High	Middle
Interview	Middle	Very good

#### 2) Banking discriminatory

Variables	Importance	Preferred type
Gender	High	Male
Age	Middle	18-34
Nationality	Low	German
Income	Middle	High
Education	Middle	Middle
Interview	Low	Good

#### 3) HR fair

Variables	Importance	Preferred type
Gender	Null	
Age	Low	18-34
Nationality	Null	
Income	Middle	High
Education	High	Middle
Interview	Low	Very good

#### 4) Banking fair

Variables	Importance	Preferred type
Gender	Null	
Age	Low	18-34
Nationality	Null	
Income	High	High
Education	Low	Middle
Interview	Low	Good

Figure 22 represents the preferences of the AI DSS graphically, depending on the sector (HR or Banking) and on whether the AI was programmed to minimize discrimination by gender and country, or not (Fair vs Discrim).

Figure 22 Preferences of the AI, by sector and type of AI



Source: Own analysis

Preferences for one dimension are presented as a combination of the importance of that dimension (Irrelevant, Low, Moderate, High), graded from 0 to 3, and the direction of the preference. The

direction was indicated by stating the most preferred characteristic (e.g. middle-aged applicants). We code the direction of preference from -1 to 1, indicating the extremities of the characteristics. For example, if asked to choose between Men and Women, -1 codes a preference for men, 1 a preference for women. If asked to state a preference in terms of the age of the applicant, 18-34-year-old are coded as -1, 35-54-year-old as 0 and 55-65-year-old as 1.

We summarize preferences by multiplying the importance of a dimension by its direction.

In Figure 22, for example, for the Discriminatory AI presented to HR professionals, we see that the most preferred gender are men, and this dimension had a high importance (+3). In terms of country, Germans are preferred but this dimension is of low importance.

The “fair” reference is shown as the middle hexagon. We see below that fair AI does not discriminate by Country and Gender. The fair and discriminatory AI do not differ much in their preferences along other dimensions in both sectors.

#### Annex 4. Preferences and prejudices of the deciders

Along with Figure 10 which showed preferences of deciders depending on their country and sector, we also show in Figure 23 the preferences of deciders by gender and country, so as to evidence homophily.

Figure 23 Deciders' preferences among applicants, by gender and country.

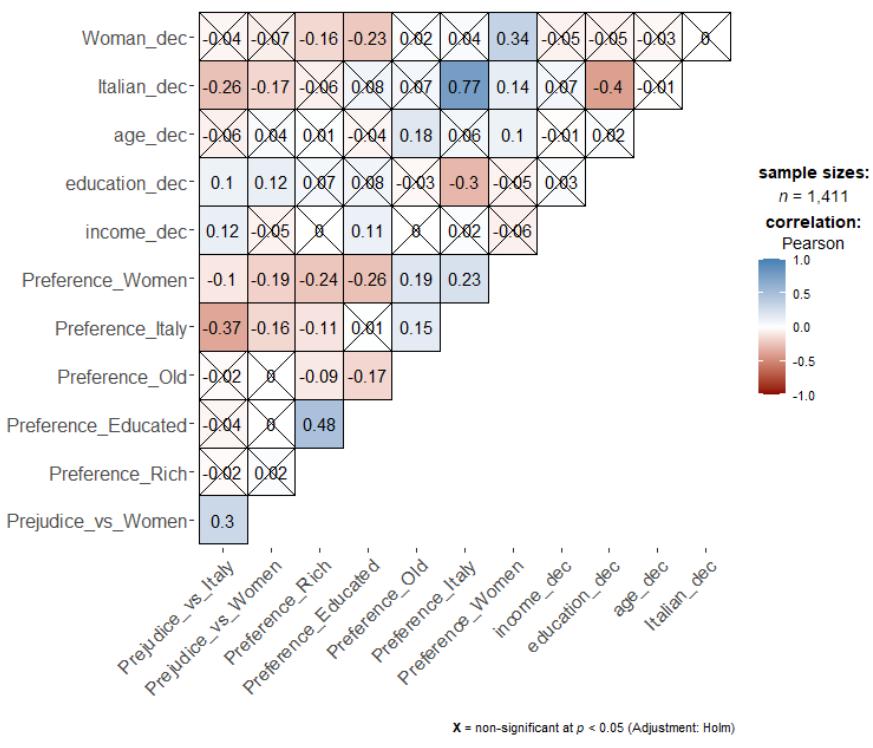


*Source: Own analysis*

We correlate preferences and prejudices of the deciders with their own characteristics (Figure 24). We find evidence of homophily in terms of nationality, gender and age, whereby Italian deciders prefer Italian applicants ( $r=0.77$ ,  $p<5\%$ ), women deciders prefer women applicants ( $r=0.34$ ,  $p<5\%$ ), and older deciders prefer older applicants ( $r=0.18$ ,  $p<5\%$ ). However, this does not extend to preferences in terms of education and income, whereby correlation coefficients are not significant.

We also find that Italian deciders have lower levels of prejudice against Italians ( $r=-0.26$ ,  $p<5\%$ ), but women deciders do not have significantly lower levels of prejudice against women ( $r=-0.07$ , n.s.).

Figure 24 Correlation between decider characteristics, preferences, and prejudices.



Source: Own analysis

## Annex 5. Regressions

Table 24 Factors in the choice among applicants, treatment without AI

	Dependent variable: choice_1				
	(1)	(2)	(3)	(4)	
W_vs_M		0.042*	0.042*	0.042*	0.045**
Ita_vs_Ger		-0.028	-0.024	-0.031*	-0.032*
education_diff		0.026**	0.026**	0.020*	0.021*
age_diff		-0.013*	-0.002	-0.003	-0.003

income_diff	0.033***	0.019**	0.014*	0.014*
interview_diff	0.032***	0.022***	0.023***	0.023***
diff_diff_applicant_gender	-0.003		-0.002	-0.002
diff_diff_applicant_country	-0.035**		-0.038**	-0.037**
diff_diff_applicant_age	0.009		0.052	0.052
diff_diff_applicant_education	-0.020		-0.022	-0.021
diff_diff_applicant_income	-0.028*		-0.037**	-0.037**
W_vs_M:Prejudice_vs_Women				-0.042
Ita_vs_Ger:Prejudice_vs_Italy				0.006
diff_diff_ideal_applicant_gender	0.00002	0.0001	0.001	
diff_diff_ideal_applicant_country	-0.007**	0.0003	0.0002	
diff_diff_ideal_applicant_age	-0.015***	-0.018***	-0.018***	
diff_diff_ideal_applicant_education	-0.003	-0.003	-0.003	
diff_diff_ideal_applicant_income	-0.013***	-0.014***	-0.014***	
diff_diff_ideal_applicant_interview	-0.014***	-0.014***	-0.014***	
Constant	0.500***	0.504***	0.503***	0.503***
Observations	5,520	5,520	5,520	5,520
R <sup>2</sup>	0.031	0.039	0.043	0.043
Adjusted R <sup>2</sup>	0.030	0.037	0.040	0.040
F Statistic	179.004***	225.004***	247.088***	249.250***

Note:

\*p<0.05; \*\*p<0.01; \*\*\*p<0.001

Source: Own analysis

Explanation of variables:

- choice\_1 is 1 if applicant 1 is chosen, 0 else (i.e. applicant 2 is chosen)
- W\_vs\_M=-(gender\_1-gender2), so it is 1 if applicant 1 is a woman and applicant 2 is a man, -1 if applicant 1 is a man and applicant 2 is a woman, 0 else
- Ita\_vs\_Ger=-(country\_1-country\_2) so it is 1 if applicant 1 is Italian and applicant 2 is German, -1 if applicant 1 is German and applicant 2 is Italian, 0 else
- education\_diff=education\_1-education\_2
- age\_diff=age\_1-age\_2
- income\_diff=income\_1-income\_2
- interview\_diff=interview\_1-interview\_2
- diff\_diff\_applicant\_gender=diff\_applicant\_1\_gender- diff\_applicant\_2\_gender

- whereby  $\text{diff\_applicant\_1\_gender} = \text{abs}(\text{gender\_dec}-\text{gender\_1})$
  - the other  $\text{diff\_diff\_applicant}$  variables are computed according to the same principle
- $\text{Prejudice\_vs\_Women} = 2 - (\text{Q7\_1} + \text{Q7\_2} + \text{Q7\_3} + \text{Q7\_4}) / 4$
- $\text{Prejudice\_vs\_Italy} = 2 - (\text{Q8\_1} + \text{Q8\_2} + \text{Q8\_3} + \text{Q8\_4}) / 4$
- $\text{diff\_diff\_ideal\_applicant\_gender} = \text{diff\_ideal\_applicant\_1\_gender} - \text{diff\_ideal\_applicant\_2\_gender}$ 
  - whereby  $\text{diff\_ideal\_applicant\_1\_gender} = \text{QRANK1\_1} * \text{abs}(\text{QRank2\_Gender} - \text{QRank2\_Gender\_1})$ .
  - whereby  $\text{QRank2\_Gender} = 1$  if the decider prefers men, -1 if the decider prefers women, 0 if he expressed no preferences and  $\text{QRank2\_Gender\_1} = 1$  if applicant 1 is a man, -1 else.
  - The other  $\text{diff\_diff\_ideal\_applicant}$  variables are computed according to the same principle.

## Getting in touch with the EU

### In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online ([european-union.europa.eu/contact-eu/meet-us\\_en](http://european-union.europa.eu/contact-eu/meet-us_en)).

### On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696,
- via the following form: [european-union.europa.eu/contact-eu/write-us\\_en](http://european-union.europa.eu/contact-eu/write-us_en).

## Finding information about the EU

### Online

Information about the European Union in all the official languages of the EU is available on the Europa website ([european-union.europa.eu](http://european-union.europa.eu)).

### EU publications

You can view or order EU publications at [op.europa.eu/en/publications](http://op.europa.eu/en/publications). Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre ([european-union.europa.eu/contact-eu/meet-us\\_en](http://european-union.europa.eu/contact-eu/meet-us_en)).

### EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex ([eur-lex.europa.eu](http://eur-lex.europa.eu)).

### EU open data

The portal [data.europa.eu](http://data.europa.eu) provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

# Science for policy

The Joint Research Centre (JRC) provides independent, evidence-based knowledge and science, supporting EU policies to positively impact society



EU Science Hub  
[Joint-research-centre.ec.europa.eu](http://Joint-research-centre.ec.europa.eu)



Publications Office  
of the European Union