ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240598

Understanding the Impact of Human Oversight on Discriminatory Outcomes in AI-Supported Decision-Making

Alexia Gaudeul[a,*], Ottla Arrigoni[a], Vicky Charisi[a], Marina Escobar-Planas[b,a] and Isabelle Hupont[a]

[a]Joint Research Centre, European Commission
[b]Universitat Politècnica de València
ORCID (Alexia Gaudeul): https://orcid.org/0000-0003-3397-4234, ORCID (Ottla Arrigoni): https://orcid.org/0009-0005-1323-1793, ORCID (Vicky Charisi): https://orcid.org/0000-0001-7677-027X, ORCID (Marina Escobar-Planas): https://orcid.org/0000-0002-4513-020X, ORCID (Isabelle Hupont): https://orcid.org/0000-0002-9811-9397

Abstract. This large-scale study assesses the impact of human oversight on countering discrimination in AI-aided decision-making for sensitive tasks. It follows a mixed method approach, including a quantitative experiment with Human Resources (HR) and banking professionals in Italy and Germany (N=1411), and qualitative analyses through interviews and workshops with participants and fair AI experts. The results show that human overseers were equally likely to follow advice from a fair AI as from a generic, discriminatory AI. Human oversight does not prevent discrimination by the generic AI. Fair AI reduces gender bias but not nationality bias. Participants' choices are neither more nor less responsive to their preferences when using an AI or when left on their own. Interviews and workshops with participants highlight individual, organizational and societal biases. In case of conflict, participants prioritize their company's interests over their own view of fairness. Participants also ask for better guidance on when to override AI recommendations. Fair AI experts stress the need for a comprehensive approach when designing oversight systems. Both technological and social aspects should be taken into consideration to ensure fairness.

1 Introduction

Artificial Intelligence (AI) aids human decision-making in high-stakes areas that can give rise to discrimination, such as credit lending and recruitment. It helps make faster decisions and avoid human cognitive biases and limitations. However, it can also introduce its own learned machine biases [19, 28, 37]. AI biases can be introduced at several stages in its development, in the same way as human biases can have different origins (Figure 1).

AI thus carries the risk of automating and perpetuating discrimination against socially marginalized groups. To deal with this issue, the General Data Protection Regulation gives the right not to be subject to a decision based solely on automated processing (Article 22) and the AI Act requires human oversight of AI systems to prevent and minimize risks to fundamental rights (Article 14). The Directive on Platform Work also addresses risks of discrimination in this way (Article 10). In this work, we investigate whether human oversight does prevent discriminatory outcomes from the use of AI. This would not be possible if overseers are subject to an automation bias and the AI is discriminatory, or if overseers are subject to algorithm aversion and thus reject recommendations from a fair AI. We show in this study that both issues are relevant, whereby overseers are influenced in the direction of discrimination suggested by a discriminatory AI, and also override suggestions made by a fair AI in order to fit their own discriminatory preferences. We draw lessons from this large-scale study about the proper way to implement human oversight to avoid discriminatory outcomes.

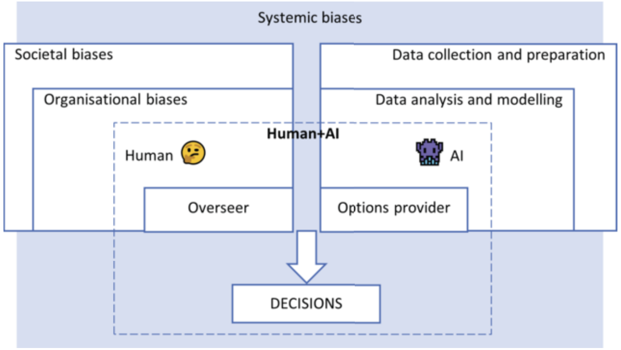


Figure 1. Human and AI decisional background and their interplay

2 Related literature

In human resources (HR), AI can be used to identify prospective recruits, check their references, or gather information about applicants from different sources [15]. In banking, AI can help determine an applicant's creditworthiness, and chatbots can help in processing loan applications and interviewing applicants [22]. The lesson from a number of scandals is that discriminatory outcomes can result from the use of such automated decision systems. This was the case for example when rating a defendant's risk of future crime [25], when deciding who is a high-risk patient needing extra healthcare [21], who gets targeted for suspension and investigation of childcare benefits [2], or for investigation of social security fraud [9]. Developers

[*] Corresponding Author. Email: Alexia.Gaudeul@ec.europa.eu.

and users of AI have worked on developing fairer AI-based decision-making processes. Much effort has been devoted to documenting and correcting bias in AI output [4, 27], and more widely, in thinking about principles for ethical AI [31, 36, 7]. Recent research has looked into whether being exposed to and getting advice from AI systems may corrupt human morals, for example, because people would accept and adopt amoral AI reasoning [18, 20, 23].

Human oversight has been proposed as an effective way to prevent negative outcomes from the use of AI. Oversight means that AI should be tested for the presence of bias in its decisions, and its decisions should be subject to review by humans. Human oversight can be classified as either ex-ante, meaning making sure that the AI is programmed correctly, or ex-post, meaning either reviewing AI suggestions before implementing them, or reviewing AI decisions if they are appealed or lead to issues [26].

In this work, we consider the interaction of ex-ante "systemic" oversight, i.e. making sure the AI is fair, and of 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. 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 biased decisions only if they maintain a willingness to question AI. Research findings about reliance on AI are quite contradictory, with a whole strand of research addressing algorithm aversion, i.e. why people tend to ignore advice from AI, while a second strand focuses on the automation bias, i.e. the tendency to follow AI advice blindly, or at least as long as it supports one's tendencies and prior belief (confirmation bias). A large behavioural literature thus documents algorithm appreciation and/or aversion, i.e. how far users follow recommendations by DSS or override them (see [29, 30, 33]).

There has been only limited research on the effectiveness of those oversight measures while taking into account the preferences of the overseer. Directly related research on the interaction of human and AI discrimination includes [3, 16, 39, 35, 1]. This research underlines selective reliance on AI depending on whether it supports preexisting beliefs. Providing biased recommendations can thus amplify preexisting bias by supporting it with apparently objective data and methods. In this paper, we further research on this topic by investigating whether even unbiased AI recommendations can be misused by discriminatory deciders, who would use unbiased AI recommendations as a baseline and thus be able to adjust them more accurately in the direction of their preferences [17].

3 Methods

We use a mixed-method research approach to examine the outcomes of a hybrid Human-AI decision process. We combine a quantitative experimental study of AI-supported decision-making with a qualitative post-experimental study based on interviews and group sessions with participants in the experiment and ideation co-design workshops with experts.

We ran an online performance-based incentivised experiment that mimicked the employer-employee and the lender-borrower relationship. The experiment went 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; and 3) asking participants in the role of "deciders" to choose among "applicants". We outline those three steps more in detail in the following.

1) 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 16th and 17th of February 2023. Participants performed a real-effort task [8] and made decisions in a trust game [6]. The real-effort task consisted in computing as many sums of four numbers as possible in five minutes. The trust game was such that applicants were lent 100 experimental currency units (ECUs), worth 4.3€, which they invested in a project that returned them 300 ECU. They were then free to send back to the lender any amount between 0 and 300 ECU.[1] We also collected participants' age, gender, country, region, level of education, nationality, occupation, sector of employment, monthly income and social class.

2) We used the data collected from the previous trust and real-effort tasks to train an AI-based Decision Support System (DSS). The objective was to construct a set of AI-based prediction models capable of assisting human deciders in subsequent decision-making experiments. To this end, we trained four Random Forest models, 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.[2] Conversely, the "generic" models included two additional sensitive or "protected" inputs: gender and nationality, which can lead to discrimination. Despite having collected a broader range of data from participants, we intentionally limited models' inputs to the attributes directly provided to deciders in subsequent experiments, ensuring that DSS assessments were grounded on the same information available to human evaluators. The models were trained with discretised input and output variables for the same reason. The original output of the trust and real-effort task were the points earned by participants, which we have transformed into a binary outcome to simulate a realistic decision scenario. This process involved setting a threshold leaving the top 30% of scores as "yes" and the remaining 70% as "no". 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 [34], employing a 5-fold cross-validation strategy over the 528 samples collected from the trust and real-effort tasks. The accuracy metrics obtained for each of the four models is shown in Table 1.

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

Table 1. Accuracy metrics for each DSS scenario.

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 tech-

[1] On average, participants managed to do 61 sums correctly in 5 minutes, and paid back 108 ECU.

[2] The interview score was based on answers to a questionnaire that evaluated an applicant's degree of motivation and self-confidence.

nique [32]. This approach aims to foster transparency and an understanding of the DSS by the deciders. Figure 2 summarizes the computational pipelines followed by the DSS to make “generic” and “fair” predictions.

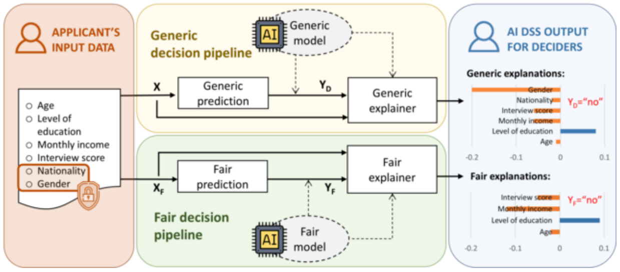


Figure 2. Pipelines followed to obtain AI-based Decision Support System’s outputs plus explanations for deciders.

The top pipeline depicts the “generic” one, where inputs to the DSS are X={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 prediction Y_D, that can be either Y_D={yes} (i.e., grant the loan or hire the applicant) or Y_D={no} (i.e., deny the loan or not to hire the applicant). Finally, the explainer, computes the explanations providing a positive/negative weight to each input variable X according to its influence on the final decision. 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={age, level of education, monthly income, interview score}. The pre-trained model used in this case to make the prediction Y_F={yes} or {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 (weight zero) the final explanations.

When developing our DSS, we took some implementation decisions we would like to discuss to be transparent with the algorithmic limitations of our DSS. Given the relatively small size of our dataset, we opted for a classic Random Forest classifier over a more complex deep learning model, as Random Forests have shown to be well-suited to achieve satisfactory performance in smaller datasets and tabular data, being less prone to overfitting [12]. On the other hand, we used cross-validation but did not evaluate on a distinct test set. The reason was that we wanted to maximize the use of our data for both training and validation. We believe this is reasonable given the dataset size, but comes with some caveats including potential overestimation of the model stability and generalization capabilities. However, 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. Regarding the “fair” decision pipeline, we initially implemented the AI Fairness 360 toolkit [5], providing bias mitigation algorithms for datasets and models. Our focus was on safeguarding against biases related to both gender and nationality. However, our trials revealed challenges in achieving satisfactory fairness metrics when protecting both variables simultaneously. Consequently, we chose to directly exclude these two inputs from the model to substantially mitigate their influence and maintain an adequate level of fairness [10]. This decision, as documented in Table 1, contributed to 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 [24]. In any case, while acknowledging these limitations, our commitment was utilizing real data and genuine machine learning models to preserve the integrity of the decision-making process, avoiding relying on synthetic data or fabricated predictions.

3) The last step was recruiting 1411 HR and banking professionals in Italy and Germany between the 24th of June and the 30th of August 2023. Participants were randomly drawn from B2B panels of HR and finance professionals based on the available profile data (occupation, age, gender and region). We had three treatments, varying whether deciders decided on their own, with recommendations of a fair AI, or with recommendations of a generic AI. Table 2 shows the distribution of deciders by country, background and treatments.

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

Table 2. Sample distribution by sector, country and treatment.

We first asked those decisions makers (DMs) for their preferences in terms of the characteristics of applicants and then asked them to choose whom to hire/lend. Participants were told they would be shown a succession of 12 pairs of applicants from whom to choose. The dimensions to judge applicants were gender, age, nationality, level of education, income and interview score. Deciders in HR were paid based on performance of one of their chosen applicant in the real-effort task. They got 4 ECU for each correct sum made by the person they hired, and the person they hired got a wage of 100 ECU (worth 4.3€). Deciders in banking were paid based on what was repaid to them in the trust game by one of their chosen applicants, who got what they kept for themselves.[3] On average, recipients received 5.8€, HR deciders received 6.0€ and banking deciders received 4.4€.

We 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). Deciders were then presented with a table recapitulating their own preferences (Table 3.

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

Table 3. Presentation of a decider’s own preferences, example.

Participants who got support from an AI DSS were told it predicted the performance of job applicants in the summing task / trust game based on their personal characteristics. We told them the

[3] Applicants were therefore not paid directly after taking part in their part of the experiment, but after DMs made their decisions whether to hire or lend to them.

generic DSS was programmed to include the impact of nationality and gender in a job applicant's grade, so relying on it may lead them to discriminate across job applicants. We told them the fair DSS was programmed to minimise the impact of those variables on a job applicant's grade, so relying on it ensured they do not discriminate across job applicants based on protected characteristics. We showed them the preferences of the AI in the same format as their own preferences (Table 3).

The main difference between fair and generic 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 generic AI. We summarize preferences of the AI 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. Figure 3 shows AI preferences for our 4 treatments (fair or generic AI, HR or banking). Both generic AIs strongly favored men, and favored Germans to a lower extent.

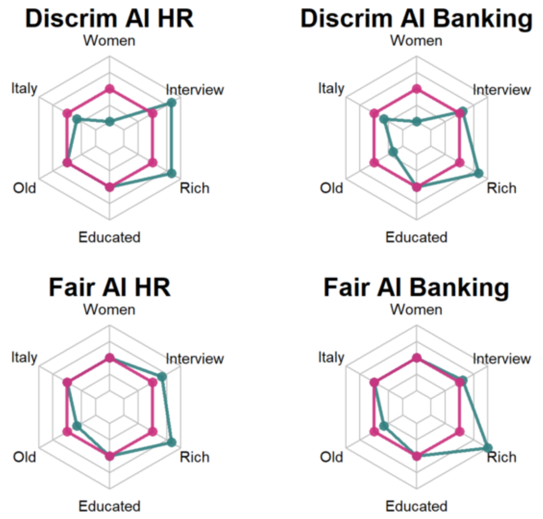
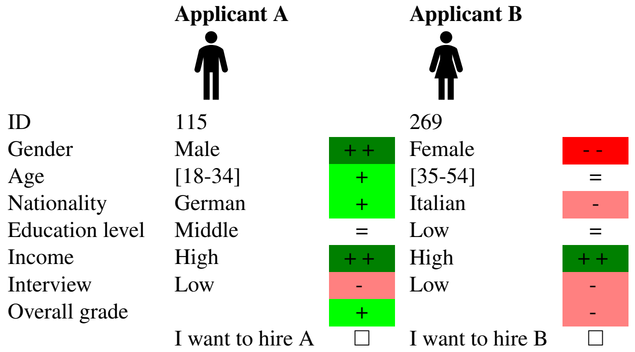


Figure 3. Preferences of the AI, by sector and fairness level.

Recommendations and decision elicitation was done as in Figure 4. In the example shown, applicant A gets a better overall grade than applicant B. Each characteristic is shown along with its grade. We see that differences in "important" characteristics result in wider differences in grades than differences in less important ones. Participants were shown a series of 12 such pairs of applicants and their choices recorded.



	Applicant A		Applicant B	
ID	115		269	
Gender	Male	+ +	Female	- -
Age	[18-34]	+	[35-54]	=
Nationality	German	+	Italian	-
Education level	Middle	=	Low	=
Income	High	+ +	High	+ +
Interview	Low	-	Low	-
Overall grade		+		-
	I want to hire A	□	I want to hire B	□

Figure 4. Decision interface including explanations for an AI's recommendation, example.

Participants were finally asked about how they made decisions, their attitudes and background in terms of 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 position, size of their company, reliance on data and DSS in their job, diversity and diversity policies in their company. We also asked participants questions about their goals, priorities, and confidence when doing the task, their perceptions of the compared honesty, work ethic, reliability and performance of men and women, and of Italians and Germans, their view on discriminating by gender or nationality, and their perception of the DSS.

We followed up the experimental study with a qualitative study: we invited volunteers from the quantitative study (N=13, defined by the method of data saturation) to participate in 1-hour interviews and 2-hours small-groups workshops. Workshops were held online in the native language of participants and were facilitated by the research team. Participants were asked to "think aloud" and reason while simulating part of the experiment and were prompted to observe different choice situations in the experiment and reflect on their own biases. In the group sessions, we discussed the results of the study and the ecological validity of the experiment. We ran a collaborative workshop in Brussels with a multidisciplinary group of fair AI researchers (N=14) to discuss the interpretation of the results from the qualitative and quantitative studies. Experts went through a speculative co-design activity about future possible interventions to tackle algorithmic discrimination in various scenarios. We concluded the study with a workshop with policymakers from across the European Commission (N=8) to discuss the results of the study and investigate the potential practical implication for existing and emerging policies.

4 Results of the experiment

The analysis in this part tests hypotheses pre-registered on the OSF registries at https://osf.io/5mz3s. The data and analytic code are at https://osf.io/mhd7r/. More detail is available in [11].

By design, the sample of deciders was balanced by gender. 70% were in the 35-54 age group, 94% had a university education, 75% had income above the median of their country, 88% worked in companies with more than 50 employees, 44% had more than 5 years' experience. They were experienced with data and statistics in their job, and 55% often used DSS.

We elicited deciders' discriminatory preferences by asking them how important each candidate characteristics were to them, and which type of candidate they favoured. Deciders rated interview, income and education most important, followed by age and nationality, and gender the least important. There were 51% of deciders who preferred male applicants, and 60% preferred Germans. Figure 5 shows preference of deciders as in Figure 3. We see a clear home bias, whereby Germans prefer German applicants, and Italians prefer Italian applicants.

We relate the choice of candidate 1 vs. candidate 2 in a pair to differences between the characteristics of the two candidates.[4] We report results from a linear probability model[5] for panel data, with

[4] The estimation equation is of the form $choice_1 = \Delta Woman + \Delta Italian + \Delta education + \Delta age + \Delta income + \Delta interview$ whereby $\Delta Woman$ is 1 if applicant 1 is a woman and applicant 2 is a man, -1 in the opposite case, 0 else. $\Delta Italian$ is computed according to the same principle. Similarly, $\Delta education = education_1 - education_2$ and so on.

[5] We can use a linear probability model whenever the relationship between probability and log odds is approximately linear over the range of mod-

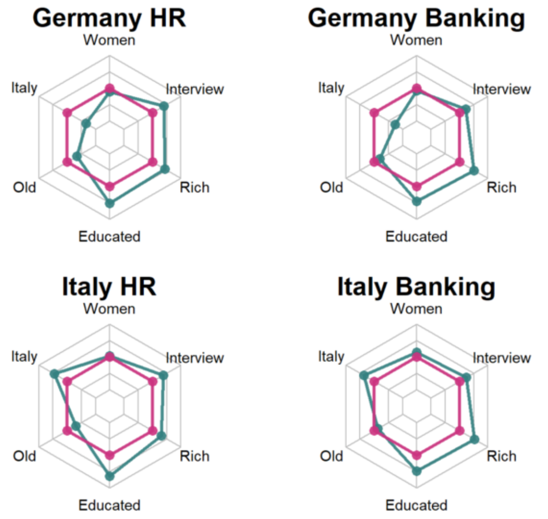


Figure 5. Preferences among applicants, by sector and country.

random individual effects (Figure 6).[6]

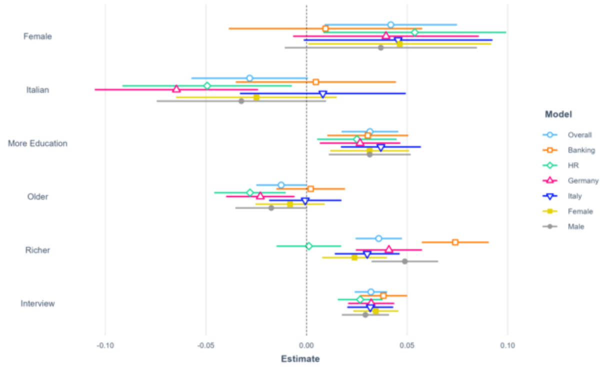


Figure 6. Impact of applicant characteristics on selection, treatment without AI.

Deciders favour of women and Germans, and favour higher levels of education, younger applicants, those with higher income and better interviews. This is consistent with their expressed preferences (Figure 5).

We now look at the extent to which deciders relied on AI recommendations in treatments with a fair and a generic AI. Full compliance with AI recommendation would be such that applicant 1 is always chosen if the difference in overall grade given by the AI is more than 0. Full compliance with both AI in both sectors was the best strategy for most participants, as this earned them significantly more than their own decisions.

Figure 7 shows the rate at which applicant 1 was chosen depending on the difference in grades between applicants. DMs choose applicant 1 only about 55% of the time even when the difference in overall grade is 2. We find they are not more likely to follow AI that is fair than AI that is discriminatory. This low rate of adherence meant that DMs did not earn more in treatments with AI than in treatments with no AI. Those findings are typical in the literature on AI reliance [13]

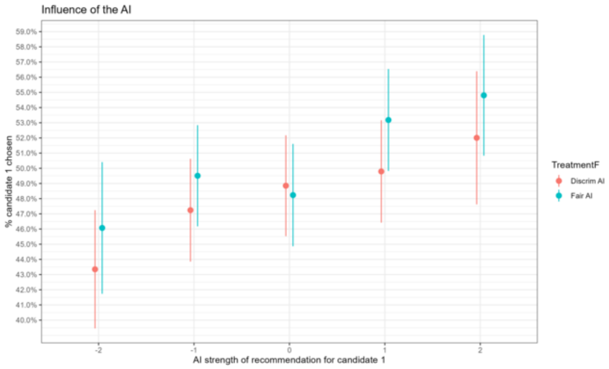


Figure 7. Likelihood to choose a candidate as a function of AI recommendation for or against that candidate.

Figure 8 shows results of regressions that confirm the positive impact of a better grade on selection of an applicant.[7] Those regressions include controls for the characteristics of applicants, rather than just the grades given to them by the AI. The AI thus has an independent effect on selection. Further regressions show that grades given by a fair AI are not more influential than those given by a generic AI.

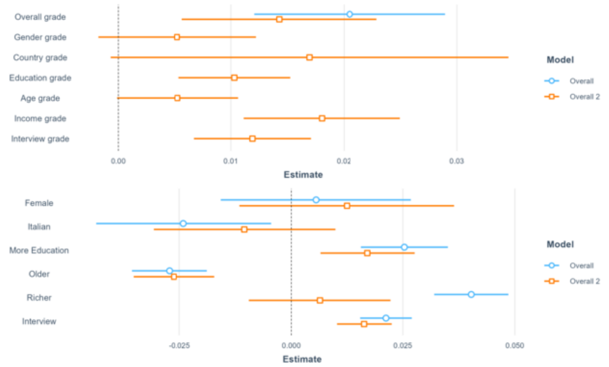


Figure 8. Influence of AI grades on choice, with controls

We now consider whether fair AI reduces discrimination and generic AI increases it (Figure 9).[8]

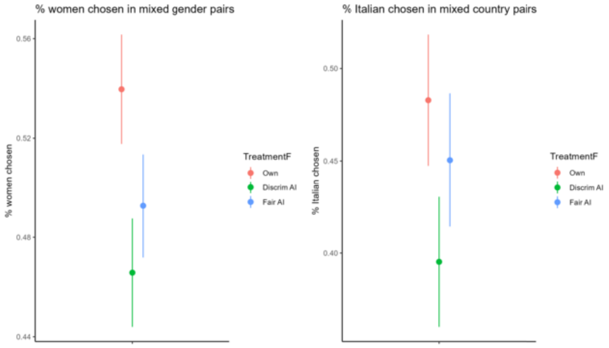


Figure 9. Gender and country discrimination, by type of AI.

elled 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.

[6] We tested the assumption of random individual effects with a Hausman-Taylor test.

[7] In addition to variables shown in footnote 4, the estimation equation includes $\Delta gender\_grade + \Delta nationality\_grade + \Delta education\_grade + \Delta age\_grade + \Delta income\_grade + \Delta interview\_grade$.

[8] The figures are based on estimating the equation in footnote 4 separately for each treatment.

We find that gender discrimination against men is disappears when using a fair AI, and changes into a discrimination against women when using a generic AI. Similarly, the generic AI results 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 show this by relating expressed preference for a type of applicant with choice of this type of applicant, depending on the AI used (Figure 10).[9]

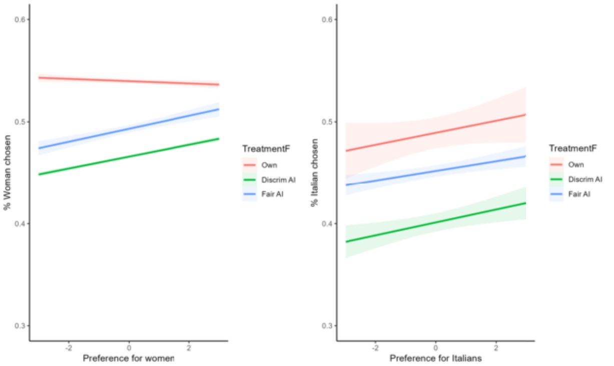


Figure 10. Bias as a function of own discriminatory preferences, for gender and nationality.

We confirm that the decider's preferences also have an impact, as the slope of the lines shown is positive. However, the magnitude of the influence of discriminatory preferences does not differ depending on the AI, as lines in Figure 10 are parallel. This means 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 (c.f. end of Section 2 on literature review).

5 Analysis of qualitative results

During the follow-up qualitative study, we organized and conducted semi-structured interviews and small-group workshops with a subset of the participants of the study, and a one-day participatory design workshop with a multidisciplinary group of experts. With the interviews and the use of open-ended questions, we focused on participants' real-life previous experience with AI, decision-making processes, perceptions of discrimination in candidates' selection and their rationale for choice in simulated scenarios from the experiment. As such, the interviews focused on three topics: a) the participant's individual and work context and current uses of AI, b) how to take account of AI advice and to react when faced with a biased or unbiased AI system, and c) priorities and biases in specific scenarios from the experiment to understand the participants' reasoning and their possible issues with the recommendations provided by the AI system. The workshops expanded on the ecological validity of the study, and gathered feedback about conceptual and methodological aspects of the experiment.

More specifically, with the qualitative study, we sought to address the following Research Questions (RQs):

- RQ1: Are professionals willing to use AI-support when making decisions, in what situations and why?
- RQ2: Do participants recognize their own biases and those in algorithms, and how do they address them?
- RQ3: What contextual factors influence how people make decisions in real life scenarios?
- RQ4: How can we envision a fairer hybrid system of AI-supported human decision-making in the experiment and in other real-life scenarios?"

Participants'responses were transcribed, annotated, and analysed using thematic content analysis. This involved cross-annotations, checks for inter-rater agreement and discussion with independent researchers to ensure reliability. The workshops were recorded for further transcription and analysis, which was considered together with the visual material produced during the workshops.

After the first iteration of the data annotation and analysis, the research team performed a second iteration taking a higher-level perspective. In this second iteration, we performed a thematic content analysis [38] based on an ad-hoc annotation scheme. The development of the first version of the annotation scheme took the Assessment List for Trustworthy Artificial Intelligence [14] as a starting point. The second version of the annotation scheme was based on the themes that emerged from the study and covered the general topics of (i) Human agency and oversight, (ii) Transparency, (iii) Diversity, non-discrimination and Fairness. Figure 11 shows the themes of the annotation scheme for each Research Question.

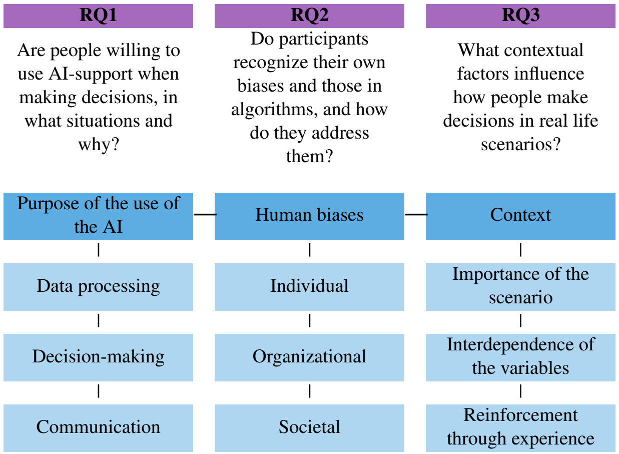

RQ1
Are people willing to use AI-support when making decisions, in what situations and why?
RQ2
Do participants recognize their own biases and those in algorithms, and how do they address them?
RQ3
What contextual factors influence how people make decisions in real life scenarios?
Purpose of the use of the AI
Human biases
Context
Data processing
Individual
Importance of the scenario
Decision-making
Organizational
Interdependence of the variables
Communication
Societal
Reinforcement through experience

Figure 11. Annotation scheme for the analysis of the interviews with the professionals.

Below, we present a summary of the outcomes of the analysis by research question. More detail is available in [11].

- RQ1 – Purpose of the use of AI: Overall, participants indicated a positive attitude toward the use of AI for their professional activities. Participants talked about their use of AI tools to support their jobs in various levels, such as for data processing, to support decision-making and to communicate with clients (e.g. chatbot).
- RQ2 – Human biases: Participants discussed the distinction between their own individual biases, those of the organisation they worked for, and broader societal prejudices. They saw their role as representing the perspective of their employer and did not express willingness to challenge organisational guidelines.
- RQ3 – Context: Participants underlined the importance of the specific use-case scenario when deciding whether to rely on AI, such

[9] We estimate equations of the form $choice_1 = \Delta Woman \times pref_{women} + \Delta Italian \times pref_{Italian} + \Delta education \times pref_{educated} + \Delta age \times pref_{old} + \Delta income \times pref_{rich} + \Delta interview \times pref_{interview}$ where preferences are computed as explained on page 3.

	Overseer		Human+AI		Decisions
Findings	Overseers favour candidates like themselves. They go along with AI discrimination if fulfilling 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 organisational biases.	Oversight of the overriding.	Mutual checks and reinforcement.	Outcome feedback and reinforcement learning.	Outcome monitoring.

Results synthesized from the qualitative study with the professionals and the reflections of the experts.

Table 4. Opportunities to improve human oversight.

as the importance of the decision (AI support was likelier to be accepted for trivial decisions), and the degree of complexity in the interactions of the different dimensions of a decision problem. They thought they were better able to judge on a case-by-case basis and to decide when and why to take into account different types of information from an applicant. Participants' willingness to rely on AI was high for analysing data, less high for guiding decisions, and lowest to elicit soft, implicit information from applicants. They thus questioned the ability of the AI to rate interviews as in our experiment. Participants also expressed their concerns about the lack of feedback on their and the AI performance, which prevented them from judging the outcome of final decisions. Finally, they underlined the need for more guidance on what to look for when deciding to override AI decisions.

During the small-group discussion regarding the ecological validity of the experiment, participants saw the relevance of the experimental set up with real-life situations, but raised issues with the selection of applicant characteristics and the grading of their importance, as well as with the AI recommendations format.

With the follow-up participatory speculative workshop, we aimed to address RQ4 in this study. The outcome of this workshop was a set of ideas and opinions about algorithms fairness and biases in AI-supported human decision-making with a view towards the future.

- RQ4 – Fair hybrid systems: The main topics that emerged were 1) Defining algorithmic and human fairness, 2) Turning the idea of fairness into practical rules and tools for human-AI collaboration, 3) Regulatory requirements for human oversight, 4) Mutual checks and questioning between human and AI, 5) Fostering users and makers (designer, developer), 6) Future policy directions.

As a final part of the study, policymakers were gathered to discuss the results of the study and examine its policy aspects. The objective was to explore the implications of our findings on human oversight and how to translate them into practical suggestions for policy makers. Groups presented their recommendations, which included proposals for new regulatory guidelines, initiatives for stakeholder engagement and training, and considerations about how to monitor and evaluate AI systems.

Findings from this research indicate a range of opportunities for improvements in oversight of AI decisions, which we summarize in Table 4. Responsible use of AI requires not only addressing the technical side (data curation, programming), or the design of better interfaces and explanations to guide decisions. It also requires guidance for human intervention and making room for exploiting synergies between humans and AI. Efforts should focus on creating oversight systems that mitigate human biases. We need to shift from relying on individuals for oversight to putting in place integrated oversight system. Promotion of systemic fairness should involve stakeholders throughout the AI life cycle to ensure that both technical and social aspects are considered. Experts and policymakers emphasized the need to assess real-world outcomes of AI-human interactions over mere rule compliance.

6 Conclusion

Our study indicates that human oversight may not prevent discriminatory outcomes from the use of AI. Rather than observing an automation bias, we found human decision makers were subject to algorithm aversion. Fair AI was overridden by deciders to favour candidates that corresponded to their own preferences, and generic AI was followed more often when its preferences corresponded to those of the deciders. Efforts to ensure non-discriminatory outcomes should therefore focus not only on programming AI that respects fairness norms, but also on putting in place *oversight systems* that ensure that users do not introduce bias in the outcomes of the AI advisory relation. The development and use of fair AI systems requires clear guidelines about when to override AI-generated recommendations. AI-assisted decisions should be regularly monitored for bias. Fairness norms at the organizational level must be reinforced with training for those overseeing AI decisions. The role of overseers must extend beyond simply approving or rejecting AI recommendations. AI systems should facilitate the inclusion of additional relevant information by the user and enable a feedback mechanism that allows overseers to contribute to the continuous improvement of the 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 also 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.

References

[1] S. Alon-Barkat and M. Busuioc. 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):153–169, 2023.

[2] Amnesty International. Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal. Available: https://www.amnesty.org/en/documents/eur35/4686/2021/en/, 2021. online.

[3] M. Avery, A. Leibbrandt, and J. Vecci. Does artificial intelligence help or hurt gender diversity? Evidence from two field experiments on recruitment in tech. *Monash Economics Working Papers*, 2023.

[4] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

[5] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, et al. AI Fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2, 2018.

[6] J. Berg, J. Dickhaut, and K. McCabe. Trust, reciprocity, and social history. *Games and economic behavior*, 10(1):122–142, 1995.

[7] V. Charisi, L. Dennis, M. Fisher, R. Lieck, A. Matthias, M. Slavkovik, J. Sombetzki, A. F. Winfield, and R. Yampolskiy. Towards moral autonomous systems. *arXiv preprint arXiv:1703.04741*, 2017.

[8] G. Charness, U. Gneezy, and A. Henderson. Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior & Organization*, 149:74–87, 2018.

[9] B. Collombat. La caisse des allocations familiales utilise un algorithme pour détecter les allocataires 'à risque'. Available: 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, 2022. online.

[10] M.-P. Fernando, F. Cèsar, N. David, and H.-O. José. Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems*, 36(7):3217–3258, 2021.

[11] A. Gaudeul, O. Arrigoni, V. Charisi, M. Escobar Planas, and I. Hupont Torres. The impact of human oversight on discrimination in AI-supported decision-making. Science-for-policy reports JRC139127, Publications Office of the European Union, Luxembourg (Luxembourg), 2024.

[12] L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520, 2022.

[13] P. Hemmer, M. Schemmer, M. Vössing, and N. Kühl. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. *PACIS 2021 Proceedings*, July 2021.

[14] High-Level Expert Group on Artificial Intelligence, European Commission. Assessment List for Trustworthy Artificial Intelligence (ALTAI), July 2020. URL https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

[15] IBM Consulting. Artificial intelligence and a new era of human resources. Available: https://www.ibm.com/blog/artificial-intelligence-and-a-new-era-of-human-resources/, 2023. online.

[16] E. Jussupow, M. A. Meza Martínez, A. Maedche, and A. Heinzl. Is this system biased?–how users react to gender bias in an explainable AI system. *42nd International Conference on Information Systems, Association for Information Systems (AIS)*, 2021.

[17] T. A. Khan. *Can Unbiased Predictive AI Amplify Bias?* Department of Economics, Queen's University, 2023.

[18] N. Köbis, J.-F. Bonnefon, and I. Rahwan. Bad machines corrupt good morals. *Nature human behaviour*, 5(6):679–685, 2021.

[19] N. Kordzadeh and M. Ghasemaghaei. Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, 31(3):388–409, 2022.

[20] S. Krügel, A. Ostermaier, and M. Uhl. Algorithms as partners in crime: A lesson in ethics by design. *Computers in Human Behavior*, 138: 107483, 2023.

[21] H. Ledford. Millions of black people affected by racial bias in healthcare algorithms. 2019. *Nature*, 574(7780):608—-609, 2019.

[22] J. Lee. The Future of AI in Lending. Available: https://www.experian.com/blogs/insights/future-ai-lending/, 2023. online.

[23] M. Leib, N. C. Köbis, R. M. Rilke, M. Hagens, and B. Irlenbusch. The corruptive force of AI-generated advice. *arXiv preprint arXiv:2102.07536*, 2021.

[24] X. Li, P. Wu, and J. Su. Accurate fairness: Improving individual fairness without trading accuracy. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, pages 14312–14320, 2023.

[25] J. A. Mattu, J. Larson, and S. Lauren Kirchner. Machine bias. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016. online.

[26] W. Maxwell. Meaningful human control to detect algorithmic errors. In C. Castets-Renard and J. Eynard, editors, *Artificial Intelligence Law: Between Sectoral Rules and Comprehensive Regime*. Larcier, 2023.

[27] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[28] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application*, 8(1):141–163, 2021.

[29] C. K. Morewedge. Preference for human, not algorithm aversion. *Trends in Cognitive Sciences*, 26(10):824–826, 2022.

[30] R. Parasuraman and D. H. Manzey. Complacency and bias in human use of automation: An attentional integration. *Human factors*, 52(3): 381–410, 2010.

[31] M. G. Reinecke, Y. Mao, M. Kunesch, E. A. Duéñez-Guzmán, J. Haas, and J. Z. Leibo. The puzzle of evaluating moral cognition in artificial agents. *Cognitive Science*, 47(8):e13315, 2023.

[32] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[33] M. Schemmer, P. Hemmer, N. Kühl, C. Benz, and G. Satzger. Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making. *arXiv preprint arXiv:2204.06916*, 2022.

[34] Scikit-Learn. Scikit-Learn library for Machine Learning in Python. Available: https://scikit-learn.org, 2024. online.

[35] F. Selten, M. Robeer, and S. Grimmelikhuijsen. 'Just like I thought': Street-level bureaucrats trust AI recommendations if they confirm their professional judgment. *Public Administration Review*, 83(2):263–278, 2023.

[36] M. Slavkovik. Mythical ethical principles for AI and how to attain them. In *ECCAI Advanced Course on Artificial Intelligence*, pages 275–303. Springer, 2021.

[37] S. Tolan. Fair and unbiased algorithmic decision making: Current state and future challenges. *arXiv preprint arXiv:1901.04730*, 2019.

[38] M. Vaismoradi, J. Jones, H. Turunen, and S. Snelgrove. Theme development in qualitative content analysis and thematic analysis. *Journal of Nursing Education and Practice*, 6(5):100, Jan. 2016. Number: 5.

[39] C. Wang, K. Wang, A. Y. Bian, R. Islam, K. N. Keya, J. Foulds, and S. Pan. When biased humans meet debiased AI: A case study in college major recommendation. *ACM Transactions on Interactive Intelligent Systems*, 13(3):1–28, 2023.