Requirements of eXplainable AI in Algorithmic Hiring

Andrea Beretta[1], Gianmario Ercoli[2], Alfonso Ferraro[2], Riccardo Guidotti[2], Andrea Iommi[2], Antonio Mastropietro[2,*], Anna Monreale[2], Daniela Rotelli[2] and Salvatore Ruggieri[2]

[1]*ISTI-CNR, Via G. Moruzzi, 1, 56124 Pisa, Italy*
[2]*Department of Computer Science, University of Pisa, Largo B. Pontecorvo, 3, 56127 Pisa, Italy*

Abstract

AI models for ranking candidates to a job position are increasingly adopted. They bring a new layer of opaqueness in the way candidates are evaluated. We present preliminary research on stakeholder analysis and requirement elicitation for designing an explainability component in AI models for ranking candidates to a job position.

1. Introduction

Employee recruiting and hiring are complex and socially-sensitive processes, whose implications have been considered from managerial, psychological, sociological, legal, and computer science perspectives. *Recruiting* covers an organization's activities to attract applicants to a job position [1]. *Hiring* (or selection) covers the activities of screening, interviewing, and selecting candidates from a given pool of applicants [2]. In support of these activities, Human Resources (HR) professionals make extensive usage[1] of Applicant Tracking Systems (ATSs). Such software tools are increasingly relying on Artificial Intelligence (AI) techniques, whose performances are intensifying the competition for human capital (the "war for talent" [3]).

In this paper, we restrict to the AI-assisted screening task of the hiring process. In fact, due to time and resource constraints, interviews can take place only for a limited number of candidates, who are selected through a preliminary screening phase. Such a screening consists of ranking candidates in the pool of applicants based on the matching of their CVs/application documents[2] with the job description, and then selecting the top candidates w.r.t. their matching score. AI models are helpful in automating the laborious, error-prone, and time-consuming task of scoring candidates. Such tools are not intended to replace HR professionals: the final judgment always rests with humans, as machines cannot accurately measure the candidate's impact, social skills, or the truthfulness of the information provided in their CVs. Nevertheless, the ranking produced by an AI model definitively influence the focus of the HR professional [4], and, ultimately, on the chance of a candidate being selected for interviewing.

AIMMES 2024 Workshop on AI bias: Measurements, Mitigation, Explanation Strategies | co-located with EU Fairness Cluster Conference 2024, Amsterdam, Netherlands
*Corresponding author.
antonio.mastropietro@di.unipi.it (A. Mastropietro)
0000-0001-8531-9325 (A. Beretta); 0000-0002-2827-7613 (R. Guidotti); 0000-0002-8823-0163 (A. Mastropietro); 0000-0001-8541-0284 (A. Monreale); 0000-0002-0943-6922 (D. Rotelli); 0000-0002-1917-6087 (S. Ruggieri)

[1]See https://www.jobscan.co/blog/fortune-500-use-applicant-tracking-systems.
[2]We will not consider here screening based on video, games, or task-based assessments.

In general, AI models can be biased in several respects [5]. This also happens for AI models used for scoring and ranking [6], and specifically in the application scenario of (algorithmic) hiring [7]. For instance, AI models can learn and replicate discriminatory patterns against protected social groups [8]. A famous case is the biased-against-women model used by Amazon [9], which was trained on the company's male-dominated workforce. Another study revealed that applicants who disclosed disabilities that would not impact their job performance received 26% less feedback compared to those who did not disclose any disability [10]. Many other works highlight the perils of employment discrimination imposed by improper design or usage of AI [11]. A strand of research applying auditing methodologies to algorithm decision-making in hiring [12] also revealed unexpected factors that affect AI prediction, such as the CV format or LinkedIn URLs [13]. The recourse to auditing instead of model's inspection is made necessary due to the proprietary nature of commercial ATSs, which prevents the disclosure of their exact internal working. However, with the increasing adoption of AI in ATSs, the opaqueness of complex and unintelligible AI models makes it impossible even for the developers of the models to understand the models' decision logic. This has boosted the research on methods to explain AI models known as *eXplainable AI* (XAI) [14].

Within the FINDHR research project (*Fairness and Intersectional Non-Discrimination in Human Recommendation*, [3]), we study XAI methods for explaining rankings of applicants to a job position in output by an AI model. Adding explanations to an AI system's output can increase users' trust [15] and mitigate bias and discrimination (see [16] specifically for hiring). From the legal side, there is an obligation for transparency in order to provide explanations about automated decision-making systems (e.g., GDPR art. 12, 13, and 14 [17]), as well as a right of candidates to contest automated decisions [18]. We are considering two approaches. First, in a *post-hoc approach*, the AI model is a given *black-box*, and we investigate ranking-specific explanations based on local methods that describe why a specific ranking was produced (*factual explanation*) and what could have changed the ranking (*counterfactual explanation*). The latter is intended to support *actionable recourse* [19]. Second, in an *explainable-by-design approach*, we design a scoring method whose internal decision logic can directly explain the produced ranking at various levels of comprehensibility. The latter approach relies on external knowledge about the hiring scenario, in the style of semantic job recommender system [20, 21].

In this paper, we discuss the preliminary phases both approaches rely on, namely *stakeholder analysis* (Section 2) and *requirements elicitation for XAI* through a multi-stakeholder *participatory design* approach (Section 3). In fact, we believe that there is the necessity of situating the problem by a thorough understanding of the stakeholders involved and of the social context where the AI models are deployed [22].

2. Hiring Process and Stakeholder Analysis

The phases of the *recruiting and hiring process* exhibit a large variability among organizations. However, they can generally be illustrated as shown in Figure 1. See [23] for a deeper discussion of the impact of AI-based tools in the process. The process starts with the publication of the job offer to attract potential candidates. The recruiting (or sourcing) activity solicits applications

[3] https://findhr.eu/

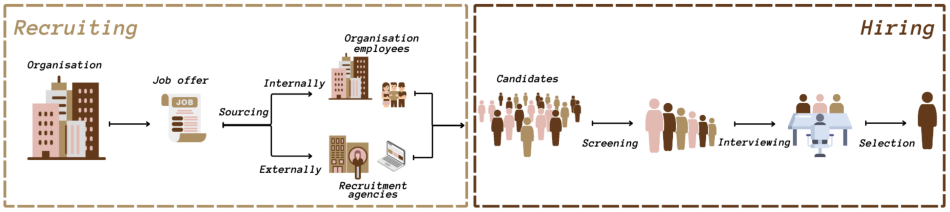

Figure 1: The recruiting and hiring process.

Table 1
Relevant internal and external stakeholders and their relation with an AI-based ATS product, and in particular with the module that ranks candidates to a job position.

Internal Stakeholders

Stakeholder	Relation to AI-based ATS	Impact
Human Resources Department (HRD)	Use the ATS to screen candidates for interview, possibly after a preliminary screening by external Recruitment Service Consultants (RSC). Interested in understanding how the ATS works, particularly on the ranking factors.	Primary
Organization Managers (OM)	Managers of the organizational unit/team with the final decision on hiring. Expect that screening by HRD/HRC will select good candidates for interviewing.	Primary
IT Department (ITD)	Responsible for integrating the ATS with the information system of the organization. May act as the ATS Software Vendors (SV) if the product is internally developed.	Primary
Current Employees (CE)	Interested in fair and consistent selection procedure across time, and in the selection of good collaborators.	Secondary
Legal and Compliance Team (LCT)	Guarantee (e.g., through impact assessment or internal auditing) that the ATS and its usage meet legal and ethical criteria for anti-discrimination, data protection, quality standards.	Secondary
Marketing or Public Relations Team (MPR)	Handle external communication, ensuring a positive image w.r.t. employee selection procedures.	Tertiary
Management Team and Executives (MTE)	Provide policy, risk mitigation, and strategic decisions (e.g., resources, job positions, etc.) for hiring. Legally responsible for the consequences of using the ATS.	Tertiary

External Stakeholders

Stakeholder	Relation to AI-based ATS	Impact
Candidates	Interact with the ATS to apply for a job position. Interested in fair, explainable, and contestable selection procedure.	Primary
Recruitment Service Consultants (RSC)	Use the ATS to screen candidates for further screening by the HRD or for interview with OM.	Primary
ATS Software Vendors (SV)	Design, develop, maintain, and run-in-the-cloud the ATS.	Primary
(Algorithmic) Auditing and Certification Agencies (ACA)	Ensure the ATS meets legal or industrial requirements w.r.t. anti-discrimination law, data protection law, quality standards, etc.	Secondary
Regulators, Policymakers, and Standardization offices (RPS)	Create laws, guidelines, and standards to ensure the ATS respects human rights and product/process quality objectives.	Tertiary
Civil Society Organizations (CSO)	Exert public scrutiny and oversight over uses of ATSs and their potential negative effects on disadvantaged people and groups.	Tertiary

from internal sources (application forms or a database of past applications) and external sources (recruiting agencies, job-focused social networks, public employment services). The pool of candidates is then scored based on the job's prerequisites, such as technical expertise, certifications, and professional background. We assume that the pool of candidates is much larger than the number of available positions, for which the screening phase is necessary and (at least, partially) automated. The top-scored candidates are invited for an interview. This phase aims to authenticate the information provided by the candidates in their CVs. Finally, the selected candidate is offered the job position.

The formalization of the hiring process helps analyzing the stakeholders involved (*stakeholder analysis*). From a *software engineering perspective* [24], stakeholder mapping charts anyone who vested interest in the ATSs, to understand their roles and inter-dependencies, their requirements, and how much the software impacts them. From a *value-sensitive design perspective*, engaging with those "who are or will be significantly implicated by the technology" [25] – beyond restricting to the users of the software – helps to establish the moral values to be taken into account at the design phase. From a *XAI perspective*, the identification of relevant stakeholders helps to tailor explanations on the basis of the identified requirements and moral values, thus dealing with constraints of different nature: cognitive (e.g., for job seekers with disability), language and culture (e.g., for foreign job seekers), technical (e.g., for HR professionals), legal (e.g., for lawyers), and abstraction-wise (e.g., explanations on single scores, on the ranking of a pool of candidates, on the overall logic of the ranking model).

Table 1 lists stakeholders that we have identified, together with the other partners of the FINDHR project, and their relation to the AI-based ATS product, in particular concerning the component that ranks candidates for a job position. We distinguish stakeholders internal to the hiring organization, and external to it. Moreover, stakeholders are categorized as primary, who have the highest influence on the process, and secondary or tertiary, who have less. Beyond stakeholder identification, we also characterize in Figure 2 the relationships among them.

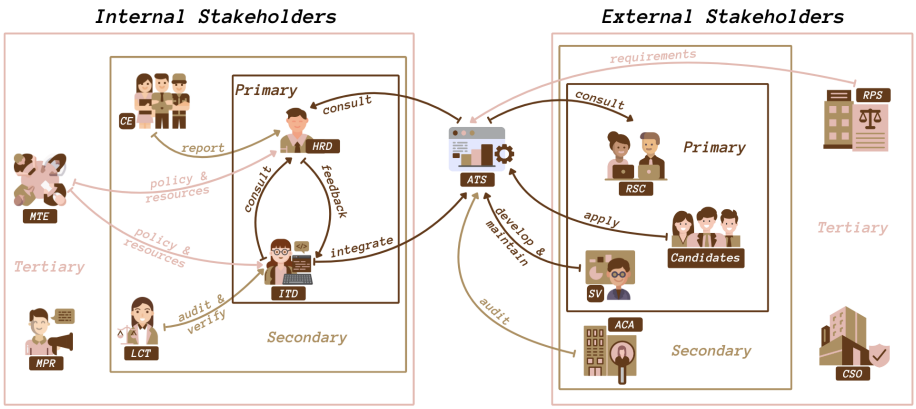

Figure 2: Relationships between relevant stakeholders.

3. Explainability Requirements for Ranking

The elicitation of explainability requirements from multiple stakeholders is being conducted in FINDHR by following a participatory design approach. The task is ongoing, for which we report in this section the planned and current activities, and the expected results.

Participatory design, namely the participation of stakeholders in the design of a (software) product, has a long tradition in software engineering [26]. Stakeholder participatory design is a much younger concept in the development of AI-based software products [27]. It targets involving the interested communities during the whole development process to prioritize AI systems that respond to human values – an objective known as *AI alignment* [28] or *socially responsible AI* [29]. Inclusion should go beyond framing the under-representation of social minorities as a data scarcity problem (a form of representation bias). Instead, it should account for preventive considerations that respond to diverse human needs and preferences. This concept is the basis for a *human-centered AI* [30]. For example, these practices can help to build socially aware language technologies that are adept for different language dialects [31]. Data-driven approaches alone are insufficient and dangerous, as data embed pre-existing biases and implicit meanings that lift to the AI model if not properly dealt with [32]. For example, candidates with different cultural backgrounds may experience issues with filling some required fields in the application form or in the CV – a signal that may mislead an AI model.

In the FINDHR project, we are organizing meetings with candidates and HR professionals. Regarding the candidates, we plan to conduct Participatory Action Research (PAR)[4], consisting of in-presence meetings (21 in 7 different European cities) with diverse legally-protected groups, including refugees and migrants, women of different ages, non-binary people, trafficked persons, young persons, women in rural and disadvantaged situations, Roma people, people with disabilities. PAR meetings will cover both fairness and explainability subjects. From the perspective of XAI requirements, we aim to gather answers to these questions (assuming the case of a candidate not being selected for an interview):

- how useful is the information about CV weaknesses and strengths of the selected candidates in redesigning future applications?
- what kind of feedback was useful in previous applications, and how it was used?
- what kind of information describing the evaluation of the CV is helpful (or demotivating)?
- what is the best way of receiving the results of an application in terms of modalities (text, graphics, etc.) and level of detail?

These questions are intended either to clarify the relevance of comparative explanations (rejected versus selected candidates) of the information conveyed in an explanation and its modality and granularity or to elucidate the expected degree of actionability in counterfactual explanations.

Regarding the HR professionals, we already interviewed six of them, working in two countries. Preliminary results of the interviews are summarized as follows:

- transparency of ranking is essential to trust and rely on it, i.e., HR professional should know what features are used and how much they weight (e.g., is experience more essential than education?), as well as which requirements the candidates meet and which they do not;

[4] https://en.wikipedia.org/wiki/Participatory_action_research

- comparative explanations based on skills or other features that distinguish candidates are useful, but they will never be followed blindly;
- understanding the score explanations might help to identify job description flaws, leading to revising the ranking (e.g., by changing weights to features or filtering candidates) and, for future job offers, to a better and fairer description.

The HR professionals we interviewed make extensive use of candidate filtering based on mandatory or preferred skills or conditions (like the distance between workplace and residence). Thus, XAI methods should adapt to these scenarios by providing explanations that dynamically change according to the active filters.

We conclude with a few design requirements for XAI in ranking candidates that emerge from the preliminary interviews and from a survey of the relevant literature on XAI methods [33–37], specifically on XAI in hiring [38–41]:

- each explanation will be tailored to a different stakeholder and their specific information needs [38, 39], with special focus on protected social groups and intersectionality issues [33];
- explanations will provide the reasons (factuals) for a given score and ranking, where scores are explained in isolation while rankings are explained by comparative approaches [39, 41];
- explanations will include actionable changes (counterfactuals) that may improve the score of candidates [39]; the actionability may be parametric in a specified cost (time or money) to implement the changes [36] or in user preferences [37];
- for score-based ranking, an interface to factuals and counterfactuals can be provided to applicants before submission, thus allowing for better personalization of the application;
- explanations will allow HR professionals to perform what-if analyses;
- explanations will have a positive impact on recruiters' trust and usage of the ATS [34];
- explanations for auditors will provide aggregate information on the past performances of the AI model, particularly w.r.t. protected social groups;
- aggregate counterfactuals can suggest public employment policies, such as prioritize training programmes for the job seekers [40];
- explanations will be fair, in the sense of an equal amount of useful information supplied to different groups of candidates [42].

4. Conclusions

We have identified the stakeholders involved in the hiring process, and their relations with an AI-based ranking of the candidates to a job position. We approach the problem of explainability of AI models for ranking by a requirement elicitation phase built on participatory design. The requirements will drive the design of both post-hoc methods and explainable-by-design ones.

Acknowledgments. Work supported by the European Union's Horizon Europe research and innovation programme for the project FINDHR (g.a. No. 101070212), and under the Excellent Science European Research Council (ERC) programme for the XAI project (g.a. No. 834756). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] J. A. Breaugh, Employee recruitment: Current knowledge and important areas for future research, Human Resource Management Review 18 (2008) 103–118. doi:10.1016/j.hrmr.2008.07.003.

[2] J. L. Farr, N. T. Tippins, Handbook of Employee Selection, 2 ed., 2017. doi:10.4324/9781315690193.

[3] J. S. Black, P. van Esch, AI-enabled recruiting in the war for talent, Business Horizons 18 (2021) 513–524. doi:10.1016/j.bushor.2021.02.015.

[4] J. M. Álvarez, A. Mastropietro, S. Ruggieri, The initial screening order problem, CoRR abs/2307.15398v2 (2024). doi:10.48550/arXiv.2307.15398.

[5] E. Ntoutsi, et al., Bias in data-driven Artificial Intelligence systems - An introductory survey, WIREs Data Mining Knowl. Discov. 10 (2020). doi:10.1002/widm.1356.

[6] M. Zehlike, K. Yang, J. Stoyanovich, Fairness in ranking, part I: score-based ranking, ACM Comput. Surv. 55 (2023) 118:1–118:36. doi:10.1145/3533379.

[7] A. Fabris, N. Baranowska, M. J. Dennis, P. Hacker, J. Saldivar, F. J. Z. Borgesius, A. J. Biega, Fairness and bias in algorithmic hiring, CoRR abs/2309.13933 (2023).

[8] E. Zschirnt, D. Ruedin, Ethnic discrimination in hiring decisions: a meta-analysis of correspondence tests 1990–2015, Journal of Ethnic and Migration Studies 42 (2016) 1115–1134. doi:10.1080/1369183X.2015.1133279.

[9] J. Dastin, Amazon scraps secret AI recruiting tool that showed bias against women, in: Ethics of Data and Analytics, Auerbach Publications, 2018. doi:10.1201/9781003278290.

[10] M. Ameri, L. Schur, M. Adya, F. S. Bentley, P. McKay, D. Kruse, The disability employment puzzle: A field experiment on employer hiring behavior, ILR Review 71 (2018) 329–364. doi:10.1177/0019793917717474.

[11] K. E. Sonderling, B. J. Kelley, L. Casimir, The promise and the peril: Artificial intelligence and employment discrimination discrimination, University of Miami Law Review University of Miami Law Revie 77 (2022) Article 3. doi:10.1002/widm.1356.

[12] M. Sloane, E. Moss, R. Chowdhury, A silicon valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability, Patterns 3 (2022) 100425. doi:10.1016/j.patter.2021.100425.

[13] A. K. Rhea, K. Markey, L. D'Arinzo, H. Schellmann, M. Sloane, P. Squires, J. Stoyanovich, Resume format, LinkedIn URLs and other unexpected influences on AI personality prediction in hiring: Results of an audit, in: AIES, ACM, 2022, pp. 572–587. doi:10.1145/3514094.3534189.

[14] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (2018) 1–42. doi:10.1145/3236009.

[15] A. Rawal, J. McCoy, D. B. Rawat, B. M. Sadler, R. S. Amant, Recent advances in trustworthy explainable Artificial Intelligence: Status, challenges, and perspectives, IEEE Trans. Artif. Intell. 3 (2022) 852–866. doi:10.1109/TAI.2021.3133846.

[16] L. Hofeditz, S. Clausen, A. Rieß, M. Mirbabaie, S. Stieglitz, Applying XAI to an AI-based system for candidate management to mitigate bias and discrimination in hiring, Electronic

Markets 32 (2022) 2207–2233. doi:10.1007/s12525-022-00600-9.
[17] European Parliament, Council of the European Union, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Official Journal of the European Union L 119 (2016). http://data.europa.eu/eli/reg/2016/679/oj.
[18] C. Henin, D. L. Métayer, Beyond explainability: Justifiability and contestability of algorithmic decision systems, AI Soc. 37 (2022) 1397–1410. doi:10.1007/s00146-021-01251-8.
[19] A. Karimi, G. Barthe, B. Schölkopf, I. Valera, A survey of algorithmic recourse: Contrastive explanations and consequential recommendations, ACM Comput. Surv. 55 (2023) 95:1–95:29. doi:10.1145/3527848.
[20] X. Chen, Y. Zhang, J. Wen, Measuring "why" in recommender systems: a comprehensive survey on the evaluation of explainable recommendation, CoRR abs/2202.06466 (2022).
[21] A. Brek, Z. Boufaïda, Semantic approaches survey for job recommender systems, in: RIF, volume 3176 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 101–111.
[22] C. Draude, G. Klumbyte, P. Lücking, P. Treusch, Situated Algorithms. A Sociotechnical Systemic Approach to Bias, Online Information Review 44 (2019) 325–342. doi:10.1108/OIR-10-2018-0332.
[23] L. Li, T. Lassiter, J. Oh, M. K. Lee, Algorithmic hiring in practice: Recruiter and hr professional's perspectives on ai use in hiring, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021, pp. 166–176. doi:10.1145/3461702.3462531.
[24] F. M. Khan, J. A. Khan, M. Assam, A. S. Almasoud, A. Abdelmaboud, M. A. M. Hamza, A comparative systematic analysis of stakeholder's identification methods in requirements elicitation, IEEE Access 10 (2022) 30982–31011. doi:10.1109/ACCESS.2022.3152073.
[25] B. Friedman, D. G. Hendry, Value Sensitive Design: Shaping Technology with Moral Imagination, The MIT Press, 2019. doi:10.7551/mitpress/7585.001.0001.
[26] M. L. Markus, J. Mao, Participation in development and implementation - updating an old, tired concept for today's IS contexts, J. Assoc. Inf. Syst. 5 (2004) 14. doi:10.17705/1jais.00057.
[27] M. Feffer, M. Skirpan, Z. Lipton, H. Heidari, From preference elicitation to participatory ML: A critical survey & guidelines for future research, in: AIES, ACM, 2023, p. 38–48. doi:10.1145/3600211.3604661.
[28] J. Ji, et al., AI alignment: A comprehensive survey, CoRR abs/2310.19852 (2023). doi:10.48550/arXiv.2310.19852.
[29] L. Cheng, K. R. Varshney, H. Liu, Socially responsible AI algorithms: Issues, purposes, and challenges, J. Artif. Intell. Res. 71 (2021) 1137–1181. doi:10.1613/jair.1.12814.
[30] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, Á. Fernández-Leal, Human-in-the-loop machine learning: A state of the art, Artif. Intell. Rev. 56 (2023) 3005–3054. doi:10.1007/s10462-022-10246-w.
[31] C. Ziems, J. Chen, C. Harris, J. Anderson, D. Yang, VALUE: understanding dialect disparity in NLU, in: ACL (1), Association for Computational Linguistics, 2022, pp. 3701–3720. doi:10.18653/v1/2022.acl-long.258.
[32] N. Shahbazi, Y. Lin, A. Asudeh, H. V. Jagadish, Representation bias in data: A survey on identification and resolution techniques, ACM Comput. Surv. 55 (2023) 293:1–293:39.

doi:10.1145/3588433.

[33] T. van Nuenen, J. M. Such, M. Coté, Intersectional experiences of unfair treatment caused by automated computational systems, Proc. ACM Hum. Comput. Interact. 6 (2022) 1–30. doi:10.1145/3555546.

[34] Y. Rong, et al., Towards human-centered explainable AI: A survey of user studies for model explanations, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024). doi:10.1109/TPAMI.2023.3331846.

[35] A. Anand, L. Lyu, M. Idahl, Y. Wang, J. Wallat, Z. Zhang, Explainable information retrieval: A survey, CoRR abs/2211.02405 (2022). doi:10.48550/arXiv.2211.02405.

[36] P. Naumann, E. Ntoutsi, Consequence-aware sequential counterfactual generation, in: ECML/PKDD (2), volume 12976 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 682–698. doi:10.1007/978-3-030-86520-7_42.

[37] J. Yetukuri, I. Hardy, Y. Liu, Towards user guided actionable recourse, in: AIES, ACM, 2023, pp. 742–751. doi:10.1145/3600211.3604708.

[38] R. Schellingerhout, F. Barile, N. Tintarev, A co-design study for multi-stakeholder job recommender system explanations, in: xAI (2), volume 1902 of *Communications in Computer and Information Science*, Springer, 2023, pp. 597–620. doi:10.1007/978-3-031-44067-0_30.

[39] M. Olckers, A. Vidler, T. Walsh, What type of explanation do rejected job applicants want? Implications for explainable AI, CoRR abs/2205.09649 (2022). doi:10.48550/arXiv.2205.09649.

[40] R. M. B. de Oliveira, S. Goethals, D. Brughmans, D. Martens, Unveiling the potential of counterfactuals explanations in employability, CoRR abs/2305.10069 (2023). doi:10.48550/arXiv.2305.10069.

[41] V. Pliatsika, J. Fonseca, T. Wang, J. Stoyanovich, ShaRP: Explaining rankings with Shapley values, CoRR abs/2401.16744 (2024). doi:10.48550/arXiv.2401.16744.

[42] N. Asher, L. de Lara, S. Paul, C. Russell, Counterfactual models for fair and adequate explanations, Mach. Learn. Knowl. Extr. 4 (2022) 316–349. doi:10.3390/make4020014.