How Data Quality Determines AI Fairness: The Case of Automated Interviewing

Lou Therese Brandner[1], Philipp Mahlow[2], Anna Wilken[2], Annika Wölke[2], Hazar Harmouch[3] and Simon David Hirsbrunner[1]

[1]*International Centre for Ethics in the Sciences and Humanities, University of Tübingen, Wilhelmstraße 19, 72074 Tübingen, Germany*
[2]*University of Cologne, Albert-Magnus-Platz, 50923 Cologne, Germany*
[3]*Hasso Plattner Institute, University of Potsdam, Prof.-Dr.-Helmert Straße 2-3, 14482 Potsdam, Germany*

Abstract

Artificial Intelligence (AI) supported job interviewing, i.e., one-sided automated applicant interviews assessed by AI-based systems, presents itself as a new mainstream solution in hiring, promising to be more efficient and effective than human recruiters, but also fairer and more objective. Selecting this technology as an illustrative case, we focus on a central element in the development of fair AI: the issue of (training) data quality (DQ). ML models with unsuitable, biased, or erroneous training data is a major source of bias in AI-based applications and therefore potentially discriminatory, unfair outcomes. However, DQ is often cast aside as one of many technical factors contributing to the overall quality of ML-based systems; this approach runs the risk of understating its crucial relevance. We select salient issues along the technology lifecycle to take a detailed look at the interrelation of fairness and DQ, illustrating how both fairness and DQ must be understood in a broad sense, taking into account normative considerations beyond technical aspects, to facilitate desirable outcomes such as the promotion of diversity, the prevention of discrimination, and the protection of workers' rights.

Keywords

Automated hiring, data quality, AI ethics, EU law

1. Introduction

Artificial Intelligence (AI) supported job interviewing presents itself as a new mainstream solution in the human resources (HR) industry. Combining machine learning (ML) techniques such as speech-to-text and text-clustering, these products and services promise to be more efficient and effective than human recruiters, but also fairer and more objective. Fairness concerns linked to the development and use of this technology have been discussed from ethical and legal perspectives, considering the increasing use of such systems in the European Union (EU) and simultaneously emerging regulatory frameworks, most famously the EU AI Act draft [9]. We argue that these discussions often neglect a central element when negotiating the development of ethical AI: the issue of (training) data quality (DQ). With a case study approach, we analyze automated job interviews to center DQ and show how other DQ dimensions directly impact – and often determine – the fairness of AI-based systems. This short article summarizes research based on the analysis of public information available about three companies offering automated interviewing – *HireVue*[2], *Knockri*[3], and *myInterview*[4] – as examples to

[2] https://www.hirevue.com, accessed 08/05/2023.
[3] https://knockri.com/, accessed 08/05/2023.
[4] https://www.myinterview.com/, accessed 08/02/2023.

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland
EMAIL: lou.brandner@uni-tuebingen.de (A. 1); p.mahlow@uni-koeln.de (A. 2); awilken2@uni-koeln.de (A. 3); awoelke@uni-koeln.de (A. 4); hazar.harmouch@hpi.de (A. 5); simon.hirsbrunner@uni-tuebingen.de (A. 6)
ORCID: 0000-0003-3827-0631 (A. 1)

assess the technology and its implications, particularly in view of the use within the EU.[5]

2. Background: Fairness in Automated Job Interviews

Automated job interviews, i.e. one-sided structured behavioral interviews with applicants that are recorded in front of a computer camera without human recruiters present, are increasingly being addressed in scientific debates [13][14][17][18][22]. Given the ubiquity of job interviews in conjunction with the time, cost, and effort that go into them, automating this process has disruptive potential for the HR industry [6]. The technology largely relies on language recognition in the shape of speech-to-text transcription and Natural Language Processing (NLP), the computational analysis of speech. Candidates are evaluated and ultimately scored or ranked for human recruiters who can base their further decisions on this automated assessment. A central claim of (semi-)automated hiring is that AI systems are less prone to bias than human recruiters, whose decisions might be unconsciously influenced by stereotypes or even consciously swayed by discriminatory behavior [7]. Given that these technical systems depend on human input and on data stemming from often discriminatory social contexts, AI applications can, however, reproduce existing biases and automate them [2][8][21]. This bears a risk to perpetuate existing job market discrimination toward women, racial minorities, and other marginalized communities [13][16][20]. The concept of discriminatory bias increasingly dominates discussions about real-world implications of AI involving data about human beings [19], with the term AI fairness or algorithmic fairness describing statistical methods intended to mitigate or eliminate these biases. Given the highly context-specific nature of fairness metrics and their dependence on different notions of fairness and societal values, historical and structural power dynamics must be taken into account when negotiating AI fairness [15].

3. Centering Data Quality

Well-established computer science DQ criteria typically include dimensions of accuracy, completeness, redundancy, readability, accessibility, consistency, usefulness, and trust [1]. Wang and Strong [23] additionally differentiate between intrinsic, contextual, representational, and accessibility aspects of datasets. Based on the growing importance of ML applications, recent attempts tailor DQ dimensions to the application field of ML [4]. Training ML models with unsuitable, biased, or erroneous training data can lead to unfair, inaccurate, and unsafe models and therefore low-quality downstream applications. The EU Agency for Fundamental Rights [10] defines training DQ as a central pillar of preventing discrimination and other unintended damages caused by AI technology. However, DQ is often viewed as one of many technical factors contributing to the overall quality of systems; this approach runs the risk of understating its essential relevance. This is all the more crucial given that the AI Act proposes several requirements for the fairness of training data for high-risk AI systems, which an interviewing system would be classified as[6]: relevance, representativity, freedom of errors, and completeness.[7] These criteria seem to be informed by the computer science literature on DQ which means they will likely become more critical for datasets that are meant to be used within the EU.

Considering the various DQ dimensions found in the literature, we analyzed information provided by the aforementioned automated interviewing platforms in the context of potential fairness concerns. Given the limited scope of this short contribution, in the following sections we will summarize a selection of salient issues to demonstrate the interrelation of DQ and fairness.

3.1 Dataset Design and Data Collection

Especially for datasets used to train models meant to assess populations with diverse socio-cultural origins and backgrounds, as is the case for hiring purposes, ensuring DQ begins before the actual data collection with the specification and documentation of use cases and target groups, for example in the

[5] HireVue is already in use in the EU by companies in Germany, France, and the Netherlands, see: https://enlyft.com/tech/products/hirevue, accessed 08/05/2023.
[6] According to Art. 6 para. 2 AIA in conjunction with Annex III no. 4 lit. a) AIA.
[7] According to Art. 10 para. 3 sen. 1.

form of datasheets [11]. Training data used for application processes should represent the anticipated language variations of the target population(s), i.e. future job applicants in the respective geographical and professional areas the system is intended to operate in. Furthermore, if training data for hiring purposes only includes certain types of industries, job roles or experience levels, the resulting system might not be able to evaluate deviating backgrounds accurately. Involving HR experts or social scientists who can research and specify target populations can thus be beneficial to avoid insufficient coverage and ensure representative, balanced, and diverse training data.

But for all analyzed companies, we observe a lack of transparency regarding the original training data; crucial DQ questions – such as the composition of datasets, when they were collected and by whom – thus remain unanswered. Available information about test datasets used to evaluate the performance of *HireVue*'s system shows a lack of people over the age of forty and makes no mention of including people with disabilities; this potentially indicates insufficient coverage of these groups which can lead to discriminatory biases and therefore an unfair treatment of individuals and groups during the assessment [2].

3.2 Labeling and Annotation

Quality flaws of labels and annotations, connected to the DQ dimensions of accuracy and objectivity, are another major source of discriminatory bias when human actors introduce their own – conscious or unconscious – social and cultural biases into the data [19], for example due to differing perceived ground truths. Automated interviewing algorithms are typically trained with pre-labeled interview data, but only provide vague information about who labels them; for instance, the *myInterview* website reads *"our machine learning models learn from our team of diverse psychologists across the world"*[8] without providing concrete information regarding specific training or demographic and geographic backgrounds.

Connected to this issue, the working conditions, employment status or salaries of labelers – a "blind spot" [12] of AI ethics which has recently started receiving more attention [24] – are not disclosed by the analyzed companies either but can significantly impact the overall labor quality and thus result in incorrect or inconsistent labels, which can introduce biases if they remain uncorrected and lower the overall reliability of the model. These factors also influence fairness in a broader sense: can a hypothetical bias-free and fairness-calibrated system truly be "fair" if those who labored for its training data were treated unethically? Can it be considered societally acceptable to use training data whose labeling has been remunerated with wages far below the EU minimum threshold?

3.3 Post-System Deployment

DQ concerns do not end with system deployment. Since biases often become apparent when systems are in active use, models require monitoring and, if biases are discovered, additional, potentially more diverse, balanced and representative training data. Systems can also continue learning during their use.[9] This on the one hand means they can potentially improve their accuracy for groups which were not sufficiently represented during training, but also that new biases can be introduced during operation, emphasizing the need for continuous monitoring and evaluation.

Particularly if the technology is used on target groups that models cannot be sufficiently trained for, rigorous post-deployment monitoring to detect limitations and offering accommodations to those who cannot be assessed fairly might present the best practice. An example in the context of automated interviewing are people with disabilities – especially those with disabilities impacting speech and thus their responses in an automated interview setting – an inherently heterogeneous group which AI-based systems currently cannot reliably recognize and categorize [5]. For instance, *HireVue* works around this issue by offering disabled candidates certain accommodations such as longer response times for interview questions and the option to directly contact the business which conducts the interview.

[8] https://www.myinterview.com/product-intelligence/, accessed 10/05/2023.

[9] Knockri, myInterview and HireVue all utilize candidate data to improve their services, which likely includes the use to further train the models: *"[…] we use certain personal data, such as User Profile Information, Video Data […] for our own purposes, such as improving and enhancing our platform and Services."*, https://www.myinterview.com/privacy/#privacy, accessed 10/05/2023.

4. Discussion

Our analysis underlines the intrinsic interconnections of DQ dimensions and fairness regarding AI-based applications in a hiring context. DQ remains relevant in all technology development stages from dataset design to post-deployment and often directly determines how bias-free and therefore fair systems can be. It is furthermore of crucial relevance regarding the legal compliance of automated job interviewing within the EU by means of the General Data Protection Regulation and anti-discrimination directives.

Transparency can be viewed as a prerequisite and bottleneck in this regard: given the lack of transparency regarding, for example, dataset composition and collection, the claimed efforts of automated hiring companies to address ethical and legal fairness concerns can only rarely be externally validated. This starts with simple issues such as a lack of information or contradictory statements on websites, for example regarding if body language is analyzed or not *(myInterview),* as well as missing resources such as reports that cannot be downloaded *(Knockri).* While *HireVue* publishes the most extensive amount of information about its systems compared to its competitors, it still misses out on detailing important DQ considerations that could considerably increase corporate accountability and public trust in the technology. Taking the described issue with modeling disability in ML systems as an example, comprehensively explaining why there is a lack of people with disabilities in training data and why systems cannot (yet) reliably and fairly evaluate these individuals could increase public understanding of ML-based technology and acceptance of alternative workarounds.

DQ and fairness must both be understood in a broad sense, taking into account normative considerations beyond technical aspects, to facilitate desirable outcomes such as the promotion of diversity, the prevention of discrimination, as well as the protection of workers' rights. Based on the insights of the study, we propose further research into scenarios beyond the discussed case of recruitment and into data synthetization, which provides both considerable opportunities and risks regarding the DQ and fairness of datasets.

5. Acknowledgements

This contribution is based on research in the KITQAR project, funded by the Policy Lab Digital, Work & Society of the German Federal Ministry of Labor and Social Affairs (BMAS).

6. References

[1] C. Batini, M. Scannapieco, Data Quality Dimensions, Springer, Cham, 2016.

[2] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities, 2019. URL: https://fairmlbook.org/.

[3] F. S. Brenner, T. M. Ortner, D. Fay, Asynchronous video interviewing as a new technology in personnel selection: The applicant's point of view, Frontiers in Psychology 7 (2016). URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2016.00863.

[4] L. Budach, M. Feuerpfeil, N. Ihde, A. Nathansen, N. Noack, H. Patzlaff, F. Naumann, H. Harmouch, The effects of data quality on machine learning performance, arXiv:2207.14529v4 [cs.DB], 2022. https://doi.org/10.48550/arxiv.2207.14529.

[5] M. Buyl, C. Cociancig, C. Frattone, N. Roekens, Tackling Algorithmic Disability Discrimination in the Hiring Process: An Ethical, Legal and Technical Analysis, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, ACM Press, New York, NY, 2022, pp. 1071–1082. https://doi.org/10.1145/3531146.3533169.

[6] T. Chamorro-Premuzic, D. Winsborough, R. A. Sherman, R. Hogan, New talent signals: Shiny new objects or a brave new world?, Industrial and Organizational Psychology 9, (2016) 621–40. https://doi.org/10.1017/iop.2016.6.

[7] T. Chamorro-Premuzic, R. Akhtar, Should companies use AI to assess job candidates? Harvard Business Review, 2019. URL: https://hbr.org/2019/05/should-companies-use-ai-to-assess-job-candidates.

[8] V. Eubanks, Automating Inequality. How High-Tech Tools Profile, Police, and Punish the Poor, St. Martin's Press, New York, NY, 2018.
[9] European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts, COM(2021) 206 final.
[10] European Union Agency for Fundamental Rights, Data Quality and Artificial Intelligence: Mitigating Bias and Error to Protect Fundamental Rights, Publications Office, Luxembourg, 2019. https://data.europa.eu/doi/10.2811/546219.
[11] T. Gebru, J. Morgenstern, B. Vecchione, J. Wortman Vaughan, H. Wallach, H. Daumé III, K. Crawford, Datasheets for Datasets, arXiv:1803.09010v8 [cs.DB], 2021. https://doi.org/10.48550/arXiv.1803.09010
[12] T. Hagendorff, Blind spots in AI ethics, AI Ethics 2 (2022) 851–867. https://doi.org/10.1007/s43681-021-00122-8.
[13] K. Houser, Can AI solve the diversity problem in the tech industry? Mitigating noise and bias in employment decision-making. 22 Stan. Tech. L. Rev. 290 (2019). URL: https://papers.ssrn.com/abstract=3344751.
[14] A. L. Hunkenschroer, C. Luetge, Ethics of AI-enabled recruiting and selection: A review and research agenda, Journal of Business Ethics 178 (2022) 977–1007. https://doi.org/10.1007/s10551-022-05049-6.
[15] J. John-Mathews, D. Cardon, C. Balagué. From reality to world. A critical perspective on AI fairness, Journal of Business Ethics 178 (2022) 945–959. https://doi.org/10.1007/s10551-022-05055-8
[16] S. Johnson, D. R. Hekman, E. T. Chan, If there's only one woman in your candidate Pool, there's statistically no chance she'll be hired, Harvard Business Review, 2016. URL: https://hbr.org/2016/04/if-theres-only-one-woman-in-your-candidate-pool-theres-statistically-no-chance-shell-be-hired.
[17] A. Köchling, S. Riazy, M. C. Wehner, K. Simbeck, Highly accurate, but still discriminatory: A fairness evaluation of algorithmic video analysis in the recruitment context. Business & Information Systems Engineering 63 (2021) 39–54. https://doi.org/10.1007/s12599-020-00673-w.
[18] 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, AIES '21, ACM Press, New York, NY, 2021, pp. 166–176. https://doi.org/10.1145/3461702.3462531.
[19] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan. A survey on bias and fairness in machine learning. ACM Comput. Surv. 54 (2022). https://doi.org/10.1145/3457607.
[20] C. Schumann, J. S. Foster, N. Mattei, J. P. Dickerson, We need fairness and explainability in algorithmic hiring, in: Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, IFAAMAS, Auckland, New Zealand, 2020, pp. 1716–1720. https://dl.acm.org/doi/abs/10.5555/3398761.3398960.
[21] A. Selbst. Disparate impact in big data policing. Georgia Law Review 52 (2017) 109–195. http://dx.doi.org/10.2139/ssrn.2819182.
[22] N. Tippins, F. Oswald, S. M. McPhail. Scientific, legal, and ethical concerns about AI-based personnel selection tools: A call to action. Personnel Assessment and Decisions 7 (2021). https://doi.org/10.25035/pad.2021.02.001.
[23] R. Wang, D. Strong, Beyond accuracy: What data quality means to data consumers. Journal of management information systems 12 (1996) 5–33. https://doi.org/10.1080/07421222.1996.11518099.
[24] A. Williams, M. Miceli, T. Gebru, The exploited labor behind artificial intelligence. Supporting transnational worker organizing should be at the center of the fight for "ethical AI", Noēma 13 (2022). URL: https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/.