Jan 17th, 12:00 AM

Proposing a Roadmap for Designing Non-Discriminatory ML Services: Preliminary Results from a Design Science Research Project

Henrik Kortum
Deutsches Forschungszentrum für Künstliche Intelligenz, Germany, henrik.kortum@dfki.de

Philipp Fukas
Universität Osnabrück, IMWI, Osnabrück, Germany, philipp.fukas@uni-osnabrueck.de

Jonas Rebstadt
Strategion GmbH, Osnabrück, Germany, jonas.rebstadt@strategion.de

Marian Eleks
Strategion GmbH, Osnabrück, Germany, marian.eleks@strategion.de

Marjan Nobakht Galehpardsari
Strategion GmbH, Osnabrück, Germany, marjan.nobakht-galehpardsari@strategion.de

See next page for additional authors

Presenter Information

Henrik Kortum, Philipp Fukas, Jonas Rebstadt, Marian Eleks, Marjan Nobakht Galehpardsari, and Oliver Thomas

Proposing a Roadmap for Designing Non-Discriminatory ML Services: Preliminary Results from a Design Science Research Project

Henrik Kortum[1, 2], Philipp Fukas[1, 2], Jonas Rebstadt[1, 2], Marian Eleks[2], Marjan Nobakht Galehpardsari[2] and Oliver Thomas[1]

[1] German Research Center for Artificial Intelligence, Osnabrück, Germany
{henrik.kortum,philipp.fukas,jonas.rebstadt,oliver.thomas}@dfki.de
[2] Strategion GmbH, Osnabrück, Germany
{henrik.kortum,philipp.fukas,jonas.rebstadt,marian.eleks,marjan.nobakht-galehpardsari}@strategion.de

Abstract. Artificial Intelligence (AI) and Machine Learning (ML) algorithms are being developed with ever higher accuracy. However, the use of ML also has its dark side. In the recent past, examples have repeatedly emerged of ML systems learning discriminatory and even racist or sexist patterns and acting accordingly. As ML systems become an integral part of both private and economic spheres of life, academia and practice must address the question of how non-discriminatory ML algorithms can be developed to benefit everyone. This is where our research in progress paper contributes. Using a real-world smart living case study, we investigated discrimination in terms of ethnicity and gender within state-of-the-art pre-trained ML models for face recognition and quantified it using an F1 metric. Building on these empirical findings as well as on the state of the scientific literature, we propose a roadmap for further research on the development of non-discriminatory ML services.

Keywords: AI, Machine Learning, Ethical AI, Non-Discrimination

1 Introduction

In recent years, Artificial Intelligence (AI) and in particular the sub-discipline of Machine Learning (ML) have gained increasing attention in research and business practice. Due to the rise of data as an essential economic resource [1] and the higher computational power available [2], ML algorithms are being developed with ever higher accuracy. However, the use of AI also has its dark side [3, 4]. Like human driven discrimination, there are cases in which ML leads to discrimination against individual groups. Algorithms used in Human Resources (HR), for example, use attributes such as place of birth for identification and thus discriminate against certain population groups, or people of color are recognized more poorly or not at all in computer vision applications [5]. Therefore, new ML inventions must meet additional requirements besides a high degree of accuracy and be tested accordingly [6]. In order to prevent such systematic discrimination, the European Union (EU) published a guideline for the

17th International Conference on Wirtschaftsinformatik,
February 2022, Nürnberg, Germany

ethical use of AI and the non-discrimination of ML algorithms [7]. Based on the EU guidelines further concrete initiatives to assess and certify the trustworthy use of AI were also developed for instance by the Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS or the Federal Office for Information Security [8, 9]. An ML algorithm producing different results for different demographic groups without these differences being professionally justified, can be seen as discriminatory. To date, there is little work on how such discrimination can be prevented in the development process of ML algorithms [10]. This paper represents the first step in an ongoing design science-orientated research approach aimed at closing this gap by answering the overarching question of how non-discriminatory ML services can be designed. Using a real-world smart living case study, we investigated discrimination in terms of ethnicity and gender within state-of-the-art pre-trained ML models for face recognition and quantified it using an F1 metric. In doing so, we also provide a practical foundation from which we derive a research roadmap for the development of non-discriminatory ML models.

2 Related Work

ML systems are increasingly used to assist humans with complicated decision-making tasks [11]. But there exist a variety of ethical and legal problems with such systems relating to transparency, accountability, explainability, and fairness [12, 13]. For example, algorithmic decision-making processes used by companies for hiring employees can lead to unfair treatment of certain groups of people, implicit discrimination and perceived injustice [10, 14]. One sample of an ML based system is face recognition, which is capable of uniquely identifying or confirming a person [15]. Despite the potential of face recognition to aid in law enforcement [16, 17] investigations, there are several significant problems with the technology. The fact that machines are often better at identifying white faces shows a persistent algorithmic bias in facial recognition technology, which may result in false positives that match a suspect's face to the incorrect identification [18]. Furthermore, even cutting-edge face recognition algorithms have been shown to be biased in terms of the input subject’s age, gender, and skin tone [19]. Developing adequate metrics to assess this bias is an important prerequisite for achieving equity in biometric systems [20–22]. Pereira et al. [23] introduced the Fairness Discrepancy Rate (FDR), which may measure recognition differences regarding different demographic groups when utilizing biometric verification systems. FDR addresses fairness by analyzing demographic disparities under the assumption of a single decision threshold [24]. Glüge et al. looked at a method for quantifying bias in a trained Convolutional Neural Network (CNN) model for face recognition. It operates by evaluating the model's "blindness" to specific facial features in face embeddings based on internal cluster validation metrics [24]. Terhörst et al. found that the famous FaceNet, has lower recognition rates for female faces compared to male faces [25]. Findings could have implications for automated face recognition systems. Drozdowski et al. looked at the issue of demographic bias in biometric systems. They discovered that demographic variables could have a substantial effect on certain biometric algorithms, and that present algorithms are biased against certain demographic groupings. They found worse biometric performance in biometric

identification systems for females and the youngest participants, as well as lower classification accuracy for dark-skinned females in the categorization of demographic characteristics from face pictures [26]. Existing public face image databases are strongly biased toward Caucasian faces, with other races (such as Latino) being considerably underrepresented. The models trained on such datasets have inconsistencies in classification accuracy, limiting the application of face analytic systems to non-white racial groups [19]. Robinson et al. built the Balanced Faces in the Wild (BFW) dataset, which balances gender and ethnic groups [27]. Inspired by the DemogPairs dataset for face images [28], the data is made up of evenly split subgroups, with an increase in subgroups, subjects per subgroup, and face pairs.

3 Research Approach

Our research sets up on the Design Science Research Paradigm and is aligned with the Information Systems Research Framework (ISRF) developed by Hevner et al. [29]. It is framed by the theoretical knowledge base and concrete practical requirements of our application domain, the smart living data ecosystem (See Figure 1). To gain an overview of the knowledge base relevant to our research, a non-systematic literature review was conducted to identify appropriate research addressing discrimination in ML algorithms. In this context, we refer to a comprehensive literature review by Köchling and Wehner [10], which we adopt for our foundation. The most important findings were briefly presented in section 2. The practical requirements result primarily from a concrete use case in the smart living domain, which is outlined as a case study in the following. The smart living data ecosystem encompasses application scenarios far beyond simple home automation and also includes other, more private areas such as smart energy management, health, elderly care or smart building security [30–33]. Thus, the domain offers diverse and promising application possibilities for ML services, while it is also characterized by strong data privacy regulations and diverse user groups that require inclusion [34]. One application that combines all these aspects is the intelligent gatekeeper. The intelligent gatekeeper is an AI service system that supports various use cases for keyless building and apartment access [32]. It involves different ML components, such as facial recognition, liveness detection, and a conversational agent, which inherently hold a risk of discrimination [35]. In a focus group interview [36] conducted with smart living experts, requirements for the gatekeeper and in particular the critical component of facial recognition were collected. All experts agree that the component must not discriminate in terms of age, gender, or ethnicity and enable equal access for all groups of residents. This requirement is the central paradigm in the implementation of the intelligent gatekeeper. In the following section, we explain the results that emerge from the case study above and present a metric for quantifying discrimination.

4 Preliminary Results of the Case Study

To assess whether and to what extent an ML algorithm is discriminating, the degree of discrimination must be made quantifiable. In the literature, this is often achieved by

assigning a metric based on differences between demographic groups (see section 2). Our research in progress takes a similar approach to Pereira and Marcel [23] and uses the F1-Score. The F1-Score is a common measure of a test's accuracy by combining precision and recall by means of the weighted harmonic mean [37]. It is chosen based on our use case, since both the precision and the recall of a facial recognition model are of importance for the smart living domain [34]. As a result, the maximum difference in F1-Score for separate demographic groups is used, as specified in equation (1):

$$\text{F1-Difference} = max\left(\left|{F_1}^{d_i} - {F_1}^{d_j}\right|\right) \forall d_i, d_j \in \mathcal{D} \quad (1)$$

Where $\mathcal{D}$ is the set of demographic groups used in the evaluation. For the evaluation, we use the BFW dataset presented in section 2. Facial embeddings are calculated for all images in the BFW dataset based on the model that is to be evaluated. Those embeddings are used to calculate distances for all face pairings in the dataset, which are then processed into labels (match/non-match) based on a threshold. Image pairs with a distance below the threshold are labeled as a match and vice versa. The threshold is fit to the face recognition model and the BFW-dataset using the optimal combination of minimized FPR with maximized TPR in the ROC-Curve of the distances. Using the classification given by the face recognition model in combination with the correct classification and labels for demographic groups from the BFW-dataset, the F1-Score is calculated for each demographic group, allowing for the calculation of the difference-metric. In our evaluation, five different face recognition models were tested, starting with the widely used Python face recognition library [38]. In addition, the well-known models VGG16 [39] and Resnet50 [40] were tested, as well as the Facenet and Openface models, which are supported by the Deepface Framework [41]. The results can be seen in the following table, displayed as the gap between the most advantaged group and the most disadvantaged group, named in the second column:

Table 1. Results of the discrimination evaluation

Face Recognition Algorithm	Affected Groups	Average F_1-Score	F_1-Difference
Python Face Recognition	White-Asian	0,84	0,188
VGG16	Male-Female	0,60	0,032
Resnet50	Black-Asian	0,93	0,033
Facenet	White-Black	0,67	0,080
Openface	Black-White	0,42	0,037

The average F1-Score was added in Table 1 to establish the overall model performance disregarding possible discrimination. The least discriminating models are VGG16 and Resnet50. VGG16 might be 2 tenths of a percentage point less discriminating, but it also performs a lot worse overall seeing as its F1-Score is 0,33 lower than Resnet50. The worst model regarding discrimination is the Python Face Recognition, which is already marked as a problem in the corresponding Wiki entry [42].

5 Discussion and Conclusion

Using a real-world smart living case study, in this paper we have investigated how discrimination occurs in ML-based face recognition. We analyzed pre-trained ML models on a dataset including faces of male and female individuals from different

ethnic groups. Using the F1 score, we could show that some models examined provide highly different detection rates in between demographic groups. These results confirm that discrimination problems can be found in common face recognition models and expand the empirical knowledge base. Since ML performance is strongly dependent on the input data used for training, testing and hyperparameter optimization, discrimination could e.g., be an issue caused by an imbalanced dataset e.g. face datasets containing underrepresented demographic groups [19]. Because features learned through face recognition models are abstract and difficult to interpret by humans, an explanation of why a model discriminates in concrete is often not possible [43]. These findings lead to the need for a systematic discrimination evaluation process created in the multi-stage procedure shown in the roadmap below. Our research directly contributes to theoretical and practical research in the fields of Information Systems, AI and ML. On the theoretical level, we provide additional empirical evidence that ML-based face recognition algorithms can lead to unintentional discrimination. Furthermore, our work highlights a research gap: For the development of ML-based face recognition algorithms in a social and ethical setting, the identified discrimination issues need to be tackled before the algorithms can be applied in the real world. On the practical level, we proposed a first approach of how ML-developers can incorporate the measurement of discrimination in their development process. We highlighted that with the practical comparison of the F1-scores between different groups, discrimination issues can be detected. Moreover, we relate the theoretical considerations directly to the practical application of ML-based face recognition based on a case study.

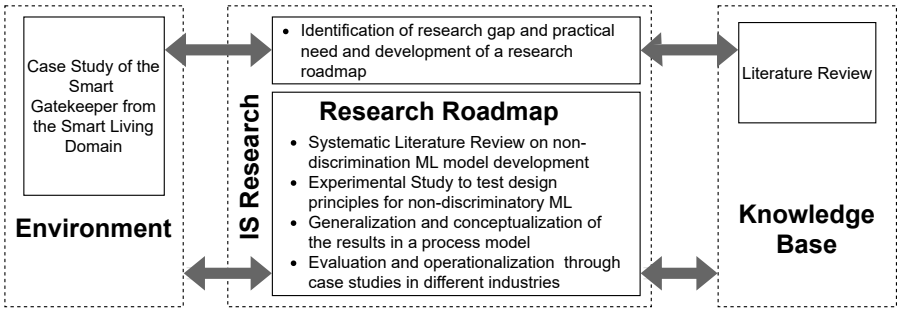

Figure 1. Research Roadmap based on the ISRF [29]

This paper presents the first step towards answering our overarching research question on how to develop non-discriminatory ML services. Based on the empirical results from the case study (section 4) and the theoretical findings (section 2), the roadmap shown in Figure 1 was derived. We want to encourage other scientists to adopt the roadmap and contribute to this field of research. The next step in our research process will be a systematic literature review to identify design principles. Since there has been limited work on how to develop non-discriminatory ML services [10], we will extend the search to other related areas where non-discrimination measures are already successfully used. After that we plan an experimental study to test and apply our findings on the use case of ML-based face recognition. We also plan to generalize and conceptualize the findings by developing a process model for the development of non-discriminatory ML algorithms. Finally, our conceptual and technical thoughts and findings will be evaluated using practical case studies.

References

1. Chen, H., Chiang, R.H.L., Storey, V.C.: Business intelligence and analytics: From big data to big impact. MIS Q. Manag. Inf. Syst. 36, 1165–1188 (2012). https://doi.org/10.2307/41703503.
2. Demchenko, Y., De Laat, C., Membrey, P.: Defining architecture components of the Big Data Ecosystem. In: 2014 International Conference on Collaboration Technologies and Systems, CTS 2014. pp. 104–112. IEEE Computer Society (2014). https://doi.org/10.1109/CTS.2014.6867550.
3. Vanderelst, D., Winfield, A.: The dark side of ethical robots. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 317–322 (2018).
4. Castillo, D., Canhoto, A.I., Said, E.: The dark side of AI-powered service interactions: Exploring the process of co-destruction from the customer perspective. Serv. Ind. J. 1–26 (2020).
5. Feuerriegel, S., Dolata, M., Schwabe, G.: Fair AI: Challenges and Opportunities. Bus. Inf. Syst. Eng. 62, 379–384 (2020).
6. Mokander, J., Floridi, L.: Ethics-based auditing to develop trustworthy AI. arXiv Prepr. arXiv2105.00002. (2021).
7. European Commission: Ethics guidelines for trustworthy AI | Shaping Europe's digital future. Brussels (2019).
8. Poretschkin, M., Schmitz, A., Akila, M., Adilova, L., Becker, D., Cremers, A.B., Hecker, D., Houben, S., Mock, M., Rosenzweig, J., Sicking, J., Schulz, E., Voss, A., Wrobel, S.: Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz - KI-Prüfkatalog. , St. Augustin (2021).
9. Federal Office for Information Security: AI Cloud Service Compliance Criteria Catalogue (AIC4). Bonn (2021).
10. Köchling, A., Wehner, M.C.: Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Bus. Res. 13, 795–848 (2020). https://doi.org/10.1007/S40685-020-00134-W.
11. de Laat, P.B.: Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? Philos. Technol. 2017 314. 31, 525–541 (2017). https://doi.org/10.1007/S13347-017-0293-Z.
12. Wachter, S., Mittelstadt, B., Floridi, L.: Transparent, explainable, and accountable AI for robotics. Sci. Robot. 2, 1–5 (2017).
13. Fukas, P., Rebstadt, J., Remark, F., Thomas, O.: Developing an Artificial Intelligence Maturity Model for Auditing. In: 29th European Conference on Information Systems Research Papers (2021).
14. Žliobaitė, I.: Measuring discrimination in algorithmic decision making. Data Min. Knowl. Discov. 2017 314. 31, 1060–1089 (2017). https://doi.org/10.1007/S10618-017-0506-1.
15. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. (2015).
16. Tol, J.: Ethical Implications of Face Recognition Tasks in Law Enforcement, https://ecp.nl/wp-content/uploads/2018/11/Artificial-Intelligence-Impact-, (2019).
17. Dushi, D.: The use of facial recognition technology in EU law enforcement:

Fundamental rights implications. Global Campus of Human Rights, Venice Lido, Italy (2020).
18. Bacchini, F., Lorusso, L.: Race, again: how face recognition technology reinforces racial discrimination. J. Information, Commun. Ethics Soc. 17, 321–335 (2019). https://doi.org/10.1108/JICES-05-2018-0050/FULL/HTML.
19. Kärkkäinen, K., Joo, J.: FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1548–1558 (2021).
20. Serna, I., Morales, A., Fierrez, J., Cebrian, M., Obradovich, N., Rahwan, I.: Algorithmic Discrimination: Formulation and Exploration in Deep Learning-based Face Biometrics. CEUR Workshop Proc. 2560, 146–152 (2019).
21. Nagpal, S., Singh, M., Singh, R., Vatsa, M.: Deep Learning for Face Recognition: Pride or Prejudiced? arXiv Prepr. 1–10 (2019).
22. Wang, M., Deng, W.: Mitigate Bias in Face Recognition using Skewness-Aware Reinforcement Learning. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 9319–9328. IEEE Computer Society (2019).
23. Pereira, T. de F., Marcel, S.: Fairness in Biometrics: a figure of merit to assess biometric verification systems, http://arxiv.org/abs/2011.02395, (2020).
24. Glüge, S., Amirian, M., Flumini, D., Stadelmann, T.: How (Not) to Measure Bias in Face Recognition Networks. In: IAPR Workshop on Artificial Neural Networks in Pattern Recognition. pp. 125–137. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58309-5_10.
25. Terhörst, P., Kolf, J.N., Huber, M., Kirchbuchner, F., Damer, N., Morales, A., Fierrez, J., Kuijper, A.: A Comprehensive Study on Face Recognition Biases Beyond Demographics. arXiv. 14, 1–14 (2021).
26. Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., Busch, C.: Demographic Bias in Biometrics: A Survey on an Emerging Challenge. IEEE Trans. Technol. Soc. 1, 89–103 (2020). https://doi.org/10.1109/TTS.2020.2992344.
27. Robinson, J.P., Livitz, G., Henon, Y., Qin, C., Fu, Y., Timoner, S.: Face Recognition: Too Bias, or Not Too Bias? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–1 (2020).
28. Hupont, I., Fernández, C.: DemogPairs: Quantifying the impact of demographic imbalance in deep face recognition. In: Proceedings - 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2019. pp. 1–7. Institute of Electrical and Electronics Engineers Inc. (2019). https://doi.org/10.1109/FG.2019.8756625.
29. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. MIS Q. 75–105 (2004).
30. Solaimani, S., Keijzer-Broers, W., Bouwman, H.: What we do – and don't – know about the Smart Home: An analysis of the Smart Home literature. Indoor Built Environ. 24, 370–383 (2015). https://doi.org/10.1177/1420326X13516350.
31. Nikayin, F., De Reuver, M.D.: What motivates small businesses for collective action in smart living industry? J. Small Bus. Enterp. Dev. 22, 320–336 (2015). https://doi.org/10.1108/JSBED-07-2012-0081.

32. Kortum, H., Gravemeier, L.S., Zarvic, N., Feld, T., Thomas, O.: Engineering of Data-Driven Service Systems for Smart Living: Application and Challenges. IFIP Adv. Inf. Commun. Technol. 592 IFIP, 291–298 (2020). https://doi.org/10.1007/978-3-030-57997-5_34.
33. Zhu, H., Samtani, S., Brown, R., Chen, H.: A deep learning approach for recognizing activity of daily living (ADL) for senior care: Exploiting interaction dependency and temporal patterns. Forthcom. MIS Q. (2020).
34. Kortum, H., Rebstadt, J., Hagen, S., Thomas, O.: Integrating Data and Service Lifecycle for Smart Service Systems Engineering: Compilation of a Lifecycle Model for the Data Ecosystem of Smart Living. In: Proceedings of the 55th Hawaii International Conference on System Sciences [accepted] (2022).
35. Foresight Newsroom, Strategion: Neues Feature für den Intelligenten Gebäudepförtner entwickelt, https://foresight-plattform.de/newsroom/neues-feature-intelligenter-interview-tuerpfoertner/, (2021).
36. Stahl, B.C., Tremblay, M.C., LeRouge, C.M.: Focus groups and critical social IS research: how the choice of method can promote emancipation of respondents and researchers. Eur. J. Inf. Syst. 20, 378–394 (2011).
37. Goutte, C., Gaussier, E.: A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In: European Conference on information retrieval. pp. 345–359. Springer Verlag (2005). https://doi.org/10.1007/978-3-540-31865-1_25.
38. Geitgey, A.: Python face-recognition, https://pypi.org/project/face-recognition/, last accessed 2020/10/09.
39. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. pp. 1–14. International Conference on Learning Representations, ICLR (2014).
40. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 770–778. IEEE Computer Society (2015).
41. Serengil, S.I., Ozpinar, A.: LightFace: A Hybrid Deep Face Recognition Framework. In: 2020 Innovations in Intelligent Systems and Applications Conference (ASYU). pp. 23–27 (2020). https://doi.org/10.1109/ASYU50717.2020.9259802.
42. Geitgey, A.: Face Recognition Accuracy Problems GitHub, https://github.com/ageitgey/face_recognition/wiki/Face-Recognition-Accuracy-Problems#question-face-recognition-works-well-with-european-individuals-but-overall-accuracy-is-lower-with-asian-individuals, last accessed 2021/08/06.
43. Williford, J.R., May, B.B., Byrne, J.: Explainable Face Recognition. In: European Conference on Computer Vision. pp. 248–263 (2020).