Article

Not peer-reviewed version

Representation Debiasing of Generated Data Involving Domain Experts

Aditya Bhattacharya * , Simone Stumpf , Katrien Verbert

Posted Date: 3 June 2024

doi: 10.20944/preprints202406.0050.v1

Keywords: explainable AI, XAI, bias detection, debiasing

Article

Representation Debiasing of Generated Data Involving Domain Experts

Aditya Bhattacharya [1,*], Simone Stumpf [2] and Katrien Verbert [1]

[1] KU Leuven Leuven, Belgium; katrien.verbert@kuleuven.be
[2] University of Glasgow Glasgow, Scotland, UK; Simone.Stumpf@glasgow.ac.uk
* Correspondence: aditya.bhattacharya@kuleuven.be)

Abstract: Biases in Artificial Intelligence (AI) or Machine Learning (ML) systems due to skewed datasets problematise the application of prediction models in practice. Representation bias is a prevalent form of bias found in the majority of datasets. This bias arises when training data inadequately represents certain segments of the data space, resulting in poor generalisation of prediction models. Despite AI practitioners employing various methods to mitigate representation bias, their effectiveness is often limited due to a lack of thorough domain knowledge. To address this limitation, this paper introduces human-in-the-loop interaction approaches for representation debiasing of generated data involving domain experts. Our work advocates for a controlled data generation process involving domain experts to effectively mitigate the effects of representation bias. We argue that domain experts can leverage their expertise to assess how representation bias affects prediction models. Moreover, our interaction approaches can facilitate domain experts in steering data augmentation algorithms to produce debiased augmented data and validate or refine the generated samples to reduce representation bias. We also discuss how these approaches can be leveraged for designing and developing user-centred AI systems to mitigate the impact of representation bias through effective collaboration between domain experts and AI.

Keywords: explainable AI; XAI; bias detection; debiasing

1. Introduction

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; **Interaction design**; • **Computing methodologies** → **Machine learning**.

The impact of Artificial Intelligence (AI) has been significant across nearly every application domain. However, the quality of the AI models largely depends on the quality of the datasets used to train them [1–5]. Moreover, several past incidents highlight the devastating consequences of using biased and erroneous datasets for training AI models, such as discriminatory treatment of users based on demographic characteristics like gender, age, race, and religion by AI systems [1,6–10]. Therefore, this increasing emphasis on data has led to a shift towards data-centric AI [11–14] approaches, where researchers and practitioners prioritise mitigating biases and issues in datasets over optimising model architectures.

Within the spectrum of biases encountered in machine learning (ML) systems [5], *representation bias* stands out as pervasive across datasets lacking a systematic data collection process [3]. Representation bias is defined as the bias in the training data that occurs due to inequality in the proportion of information leading to models that are biased in favour of the majority samples [3,5]. For instance, a dataset captured through a survey to measure screen time on electronic gadgets for teenagers could have representation bias if only high school students are involved and if home-schooled students or dropouts are not involved in the survey. Consequently, if the dataset is not representative of an entire population, the outcome of prediction models used in decision systems may not be trustworthy for the sub-populations that are under-represented. This is because the model has fewer data points to learn from for such under-represented populations, and hence, it does not generalise well for such sub-populations.

 doi:10.20944/preprints202406.0050.v1

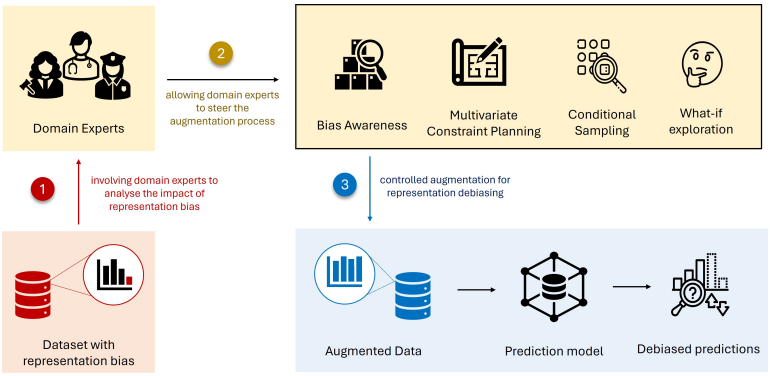

Figure 1. This paper presents the interaction approaches for representation debiasing of generated data, highlighting the crucial role of domain experts in steering data augmentation methods to mitigate representation bias within datasets.

The most effective practical solution to mitigate representation bias is through *data augmentation* [3,15,16]. Data augmentation is a technique used to increase the size and diversity of training datasets by generating new samples while ensuring that the statistical properties of the original dataset are preserved [3,17]. For instance, if a particular group is underrepresented in the original dataset, data augmentation can be used to generate more synthetic samples of that group. This method essentially allows the prediction model to learn from more samples to avoid overfitting or underfitting problems [17]. By generating more samples of underrepresented groups, the resulting augmented dataset becomes more balanced and representative of the real-world population, mitigating the possibility of bias [3].

However, an uncontrolled data augmentation, i.e., the process of applying augmentation algorithms without strict constraints or predefined rules, can still retain the representation bias in the data as these methods try to preserve the distribution and the statistical properties of the original dataset [18**?** –20]. Additionally, data augmentation using up-sampling methods like SMOTE [21] has been criticised for increasing data issues like data drift, correlated features, duplicates, or even outliers [18,22].

Therefore, domain knowledge is important for applying constraints to avoid the generation of problematic samples. It is also necessary for identifying, modifying and removing synthetic samples that may add more bias. Thus, domain knowledge can be crucial in the debiasing process of representation bias, highlighting the importance of involving domain experts in the process of representation debiasing in ML systems.

To address the current limitations in mitigating representation bias, this paper introduces our interaction approaches for representation debiasing of generated data by involving domain experts to steer data augmentation methods. These approaches aim to facilitate domain experts in explaining the presence of representation bias in datasets and leverage their expertise to generate unbiased augmented samples. Thus, interaction systems offering these interaction approaches can reduce representation bias in their training datasets and consequently, improve the overall prediction models and the quality of the training datasets for unbiased predictions. Furthermore, through an exploratory user study, we have instantiated these approaches into designs of visual components for the healthcare domain. Therefore, these interaction approaches can take us one step closer towards achieving debiased user-centred AI systems.

2. Background and Related Work

2.1. Representation Bias

As discussed earlier, representation bias arises due to the absence of sufficient information for the various subgroups or sub-categories in the dataset [3]. This type of bias can be introduced from a

 doi:10.20944/preprints202406.0050.v1

number of factors such as historical trends observed in past events, skewness in the distribution of the data, procedures used for data preparation and acquisition, and selection and sampling biases [3,5]. Representation bias is also one of the main concerns for achieving group or sub-group fairness, as models trained on data lacking representation from certain groups or subgroups are expected to exhibit lower accuracy when applied to these under-represented categories [3]. Thus, training datasets must include sufficient samples from the "less popular segments" of the data space in order for the ML system to handle these segments well. Yet, representation bias is considered to be one of the critical problems in ML systems, leading to biased and unfair predictions [3].

Representation bias is measured through two main metrics: (1) Representation rate and (2) Data coverage [3]. Representation rate of a sub-group is defined as the ratio of its sample counts to the highest frequency of occurrence among all sub-groups. Suppose if variable A has three sub-groups: x, y and z, such that $x + y + z = N$, where N is the total number of samples for variable A. Then, the representation rate of sub-group x is measured by: $r_x = \frac{x}{max(x,y,z)}$. For instance, let us consider a variable representing education level of job applicants that has three sub-categories: high-school degree, bachelor's degree and master's degree. Suppose that the dataset has 600 samples such that the number of applicants with education level as high-school degree is 100, bachelor's degree is 300 and master's degree is 200. Then, the representation rate of applicants with a high-school degree is: $r_1 = \frac{100}{max(100,300,200)} = 0.33$. Similarly, the representation rates of the other two sub-groups are 1.0 and 0.67, respectively. This implies that applicants with a high school degree are less represented than applicants with bachelor's or master's degrees in the dataset.

Meanwhile, data coverage is defined as the minimum number of samples that should be present for each sub-category. For instance, if the data coverage is set as 150, then the sub-category of applicants with an education level of a high-school degree does not meet the data coverage criteria as it has only 100 samples. Regardless of the data space, it is critical to have a high enough coverage for all significant sub-populations in the data to ensure their adequate representation.

2.2. Data Augmentation

To address representation bias, AI practitioners have recommended improving the data collection process and gathering additional data for under-represented segments of the dataset [3,17]. Nonetheless, due to practical limitations and challenges, collecting new data may not always be feasible [17]. Thus, researchers have proposed generating synthetic samples for the under-represented segments as a debiasing approach for representation bias [3,23]. Data augmentation is the process of generating new data points by applying various transformations or modifications to the existing data [3,17,19]. AI practitioners have also relied on data augmentation methods to address the class imbalance problem, which arises when the target variable predominantly contains samples from one of its sub-categories [24]. Methods such as SMOTE [21] and ADASYN [25] have been commonly used in such scenarios as they can identify the majority and the minority categories and generate samples for the minority category for creating a balanced dataset.

Despite the significance of such up-sampling techniques, these methods have been criticised for introducing data quality issues such as data drift, duplicate records and outliers [18,20,22]. Additionally, many generative AI algorithms, such as GANs [26] and VAEs [27], have also been explored by practitioners for generating synthetic samples of the data. For tabular datasets, their different variants, such as CTGAN and TVAE [28], have been considered better as they introduce fewer data issues. By creating more samples of underrepresented groups, these methods can produce an augmented dataset that is more balanced and representative of the real-world population, reducing the likelihood of bias in prediction models [29].

 doi:10.20944/preprints202406.0050.v1

However, researchers have also noted that an uncontrolled data augmentation process can impact prediction models by generating practically implausible samples [17,24,30]. For example, let us consider an under-represented diabetes prediction dataset that has factors such as glucose, body mass index (BMI), age and obesity level of a patient. An uncontrolled data augmentation process can generate samples where the obesity level is high, yet the BMI is less than 18. But practically speaking, a patient with a BMI less than 18 is considered to be underweight and not obese. Moreover, uncontrolled data augmentation can retain the existing representation bias in the data in an attempt to preserve the statistical properties of the data [20,31**?**].

Domain knowledge is considered crucial for performing controlled data augmentation for representation debiasing as it involves setting constraints to prevent the generation of problematic samples. Our work argues the importance of involving domain experts in the process of representation debiasing. Establishing an effective collaboration between domain experts and AI systems is essential for representation debiasing of generated data.

2.3. Model Steering Involving Domain Experts

Researchers working in interactive machine learning (IML) have primarily focused on investigating various approaches for user-in-the-loop model steering to improve prediction models [32–37]. Previous studies have also examined diverse approaches through which end-users could guide ML systems and actively participate in training, fine-tuning, and debugging prediction models [32–34,38–42].

Recent works have shown the importance of the active involvement of domain experts in model steering as they can easily identify and correct training data issues that are essential for steering models [37,43,44]. It has also been argued that for identifying various types of biases in the data a thorough domain knowledge possessed by domain experts is extremely important [1,45]. For instance, rich domain knowledge possessed by medical experts facilitates them in investigating the health records of patients and identifying biases or erroneous data that can lead to biased models.

Although recent works have designed and developed innovative approaches for enabling domain experts to steer prediction models [35–37,44], prior research has largely overlooked their involvement in the process of generating data through controlled data augmentation. Specifically, leveraging domain experts' prior knowledge presents significant potential for mitigating the risks associated with uncontrolled data augmentation and addressing representation bias [30]. Thus, our work aims to fill this gap and explore how domain experts can be involved in the process of representation debiasing of generated data.

 doi:10.20944/preprints202406.0050.v1

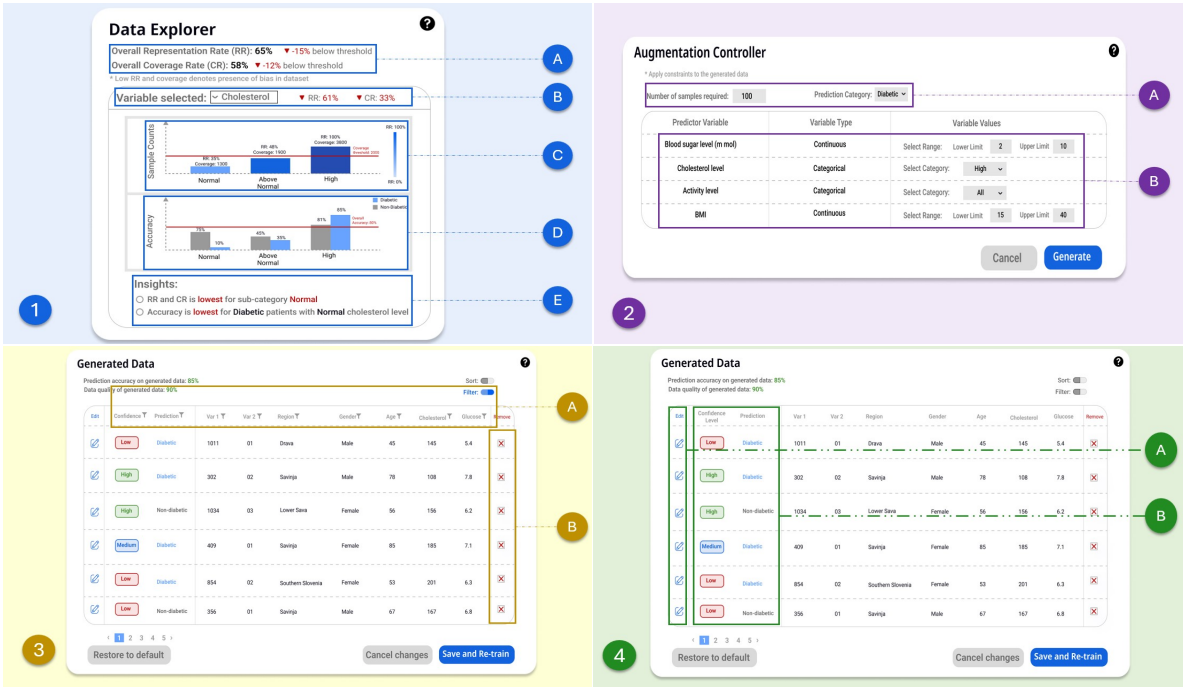

Figure 2. Design of visual components of a low fidelity prototype instantiating the interaction approaches for representation debiasing.
1 Bias awareness involves: A presenting the overall representation bias measures, B presenting the representation bias measures for each predictor variable, C showing the representation bias for each sub-category of the selected predicted variable, D showing the corresponding impact on the model performance, and E highlighting the most impacted variables and sub-categories.
2 Multivariate constraint mapping involves: A allowing users to specify the required number of generated samples for a specific target class, and B set constraints on the predictor variable values.
3 Conditional sampling involves: A allowing users to sample generated data through conditions defined in data filers, and B remove misfit samples.
4 What-if exploration involves: A allowing users to validate or modify generated samples through "what-if" analysis, and B applying the prediction model on generated samples to identify problematic samples having low prediction confidence level.

3. Interaction Approaches for Representation Debiasing

This section presents our interaction approaches for involving domain experts in the representation debiasing of training data used in ML systems. We elaborate on the rationale, description and purpose of these approaches in the following part. We have also conducted an exploratory user study (discussed in Section 4) to instantiate these interaction approaches for representation biasing through visual components of a low-fidelity prototype, as illustrated in Figure 2. The proposed interaction approaches are as follows:

- Bias awareness: Although domain knowledge is essential to identify representation bias, prior researchers have signified the importance of interactive explanations for assisting domain experts in elucidating the behaviour of prediction systems [35,37,46,47]. Thus, we propose an approach for *bias awareness*, that aims at guiding domain experts to identify biased predictor variables with the help of data-centric explanations [8,46,48]. We recommend allowing domain experts to explore the distributions of categories or sub-categories of predictor variables, including the representation rate and data coverage of each category or sub-category, through interactive visualisations for higher transparency. This interaction approach is aligned with the ML transparency and exploration principles from Bove et al. [49], which aims at providing better contextualised explanations for the presence of representation bias. Furthermore, we propose illustrating the model performance corresponding to each category or sub-category of predictor variables to identify those most impacted by representation bias.
- Multi-variate constraint planning: Despite generating additional samples of underrepresented data, one primary reason for the limited effectiveness of data augmentation

 doi:10.20944/preprints202406.0050.v1

algorithms in mitigating representation bias is the generation of practically infeasible samples [17,31]. This occurs because data augmentation algorithms typically treat each predictor variable independently rather than jointly. It requires in-depth domain knowledge to understand the joint impact of the predictor variables. With *multivariate constraint planning*, we propose empowering domain experts to impose constraints on multiple predictor variables. This allows for the generation of specific sets of samples considered essential by experts to mitigate the impact of representation bias. For example, consider the representation debiasing of a diabetes prediction dataset. If healthcare experts identify that only 50 samples of diabetic patients aged 50 to 60 with high cholesterol and high blood pressure are necessary, then multivariate constraint planning can be utilised to achieve their requirement. This interaction approach enables control over the data augmentation process to mitigate the issue of generating practically infeasible data points.

- CONDITIONAL SAMPLING: The interaction approach of *conditional sampling* is applicable after the application of data augmentation algorithms, allowing domain experts to select only relevant synthetic samples for upsampling the original training data. We recommend providing data filters for setting conditions defined by domain experts, allowing them to identify and remove generated samples which appear to be a misfit. This approach aims at purifying the generated data for minimal introduction of problematic data points during representation debiasing.
- WHAT-IF EXPLORATION: The interaction approach for *what-if exploration* of the generated data further allows domain experts to validate the generated samples. It aligns with the concept of "what-if" explanations [8,46,50,51], aiming to enhance understanding of the generated samples and their potential impact on prediction models. We recommend applying the prediction model to each generated sample to obtain their predicted target class and the corresponding confidence levels. This can allow domain experts to identify problematic data points that are difficult to train by the prediction algorithm. Additionally, domain experts should be able to adjust the values of generated samples and conduct what-if analysis to rectify such problematic generated instances.

4. Exploratory User Study

An exploratory user study was conducted with 5 healthcare experts (2 females, 3 males; age: 29 - 51 years), each having more than four years of experience in healthcare, from the Faculty of Healthcare Sciences, University of Maribor, Slovenia, to explore the design space of UI components offering our proposed interaction approaches for representation debiasing. The study included individual co-design and think-aloud sessions (each session lasted for about 30 minutes on average). The sessions were recorded and transcribed for analysing the feedback of the participants. The goal of this study was to explore the design space of UI components for the interaction approaches for representation debiasing.

The findings from this exploratory study allowed us to design visual components for a low-fidelity click-through prototype. The design of the visual components is shown in Figure 2. In this study, we extensively engaged healthcare experts to gather their perspectives on participating in the task of mitigating representation biases within ML systems. Encouragingly, all participants expressed enthusiastic support for actively contributing to the debiasing process. Additionally, they have mentioned that this process could allow them to better understand predicted outcomes that do not match their expectations. Therefore, as a key takeaway of this study, we came up with the following UI components supporting the four interaction approaches discussed in Section 3:

1. DATA EXPLORER - This component is designed creating *bias awareness* (Figure 2 (1)). It includes presenting the overall representation bias measures, as well as the representation bias measures for each predictor variable. We included a visual representation of the distribution of each predictor variable, illustrating the representation bias for each sub-category of the selected variable. Additionally, we illustrated the corresponding impact on the model performance for each sub-category and highlighted the most impacted variables and sub-categories.

 doi:10.20944/preprints202406.0050.v1

2. AUGMENTATION CONTROLLER - This component is designed to support *multivariate constraint mapping* (Figure 2 (2)). It is designed to allow users to specify the required number of generated samples for a specific target class, and set constraints on the predictor variable values for the generated samples.
3. GENERATED DATA EXPLORER - This component is designed to support *conditional sampling* (Figure 2 (3)) and *what-if exploration* (Figure 2 (4)) of generated data. It is designed to allow users to sample generated data through conditions defined in data filers, and remove problematic samples. Moreover, it allows users to validate or modify generated samples through "what-if" analysis. By applying the prediction model to generated samples, this component further facilitates users to identify problematic samples having low prediction confidence levels.

5. Discussions

5.1. Debiasing is a Continual Process

As noted by prior researchers [3**?**], debiasing ML systems is a continual process as new sources of bias can get introduced when the system is deployed in different contexts or deployed with new data (including generated data). Although findings from our exploratory study indicate that domain experts can be involved in representation debiasing, their continuous involvement in the debiasing process could be a challenge due to their limited availability. Yet, continuous monitoring and evaluation of prediction models are necessary to ensure that these models remain unbiased and fair. Further investigation is required to analyse how effectively can user-centric AI systems support continuous involvement of domain experts in the debiasing process.

5.2. Implications on Fairness

The presence of representation bias is one of the primary concerns for group fairness or sub-group fairness of AI systems [3,5,45]. Traditional approaches for mitigating fairness issues involve identifying and removing protected attributes such as gender, race, religion, etc., from the training data for the prediction models [45]. However, only eliminating a protected attribute may not be sufficient if the model has access to other predictor variables that are proxies for the protected attribute [**?**]. For instance, if gender is removed as it could be a protected attribute, other attributes such as height, weight or BMI can act as proxy attributes. However, the interaction approaches proposed in this paper can provide a solution to the limitations of these traditional approaches for achieving group or sub-group fairness. Thus, we hypothesise that these interactions could take us closer towards achieving group or sub-group fairness in AI.

5.3. Future Work

Future research should consider applying the visual components of representation debiasing to design and develop user-centred AI systems tailored for domain experts. More extensive user studies should be conducted to investigate how these approaches can empower domain experts to improve the overall prediction model performance and data quality of AI systems. Furthermore, it will be interesting to analyse how the process of representation debiasing affects the trust and understanding of domain experts on AI systems.

6. Conclusions

To conclude, this paper introduces our proposed interaction approaches for representation debiasing of generated data used in AI systems by allowing domain experts to steer the data augmentation process. In this paper, we describe these approaches, elaborate on the rationale behind them and also present our design of UI components to instantiate these interaction approaches for user-centred AI systems tailored for domain experts. We aim to discuss the implications of these UI components for human-in-the-loop AI systems to mitigate representation bias in the workshop and

 doi:10.20944/preprints202406.0050.v1

gain valuable feedback for proposing a framework for representation debiasing, enabling user-centred AI practitioners to take a step closer towards achieving debiased systems.

Acknowledgments: We thank Maxwell Szymanski and Robin De Croon for their valuable feedback on this research. This research was supported by Research Foundation–Flanders (FWO grants G0A4923N and G067721N) and KU Leuven Internal Funds (grant C14/21/072) [52]. We also thank our participants from the University of Maribor for their feedback captured from the exploratory user study.

References

1. Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights. 2019.
2. Ding, J.; Li, X. An Approach for Validating Quality of Datasets for Machine Learning. 2018, pp. 2795–2803. doi:10.1109/BigData.2018.8622640.
3. Shahbazi, N.; Lin, Y.; Asudeh, A.; Jagadish, H.V. Representation Bias in Data: A Survey on Identification and Resolution Techniques. *ACM Comput. Surv.* **2023**, *55*. doi:10.1145/3588433.
4. Aldoseri, A.; Al-Khalifa, K.N.; Hamouda, A.M. Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges. *Applied Sciences* **2023**, *13*, 7082. doi:10.3390/app13127082.
5. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A Survey on Bias and Fairness in Machine Learning, 2022, [arXiv:cs.LG/1908.09635].
6. Khan, B.; Fatima, H.; Qureshi, A.; Kumar, S.; Hanan, A.; Hussain, J.; Abdullah, S. Drawbacks of Artificial Intelligence and Their Potential Solutions in the Healthcare Sector. *Biomedical Materials & Devices* **2023**, pp. 1–8. Advance online publication, doi:10.1007/s44174-023-00063-2.
7. Ahmad, S.; Han, H.; Alam, M.e.a. Impact of Artificial Intelligence on Human Loss in Decision Making, Laziness and Safety in Education. *Humanities & Social Sciences Communications* **2023**, *10*, 311. doi:10.1057/s41599-023-01787-8.
8. Bhattacharya, A. Applied Machine Learning Explainability Techniques. In *Applied Machine Learning Explainability Techniques*; Packt Publishing: Birmingham, UK, 2022.
9. Armstrong, S.; Sotala, K.; Ó hÉigeartaigh, S.S. The Errors, Insights and Lessons of Famous AI Predictions – and What They Mean for the Future. *Journal of Experimental & Theoretical Artificial Intelligence* **2014**, *26*, 317–342. doi:10.1080/0952813X.2014.895105.
10. Papagiannidis, E.; Mikalef, P.; Conboy, K.; van de Wetering, R. Uncovering the dark side of AI-based decision-making: A case study in a B2B context. *Industrial Marketing Management* **2023**, *115*, 253–265. doi:10.1016/j.indmarman.2023.10.003.
11. Mazumder, M.; Banbury, C.; Yao, X.; Karlaš, B.; Rojas, W.G.; Diamos, S.; Diamos, G.; He, L.; Parrish, A.; Kirk, H.R.; Quaye, J.; Rastogi, C.; Kiela, D.; Jurado, D.; Kanter, D.; Mosquera, R.; Ciro, J.; Aroyo, L.; Acun, B.; Chen, L.; Raje, M.S.; Bartolo, M.; Eyuboglu, S.; Ghorbani, A.; Goodman, E.; Inel, O.; Kane, T.; Kirkpatrick, C.R.; Kuo, T.S.; Mueller, J.; Thrush, T.; Vanschoren, J.; Warren, M.; Williams, A.; Yeung, S.; Ardalani, N.; Paritosh, P.; Zhang, C.; Zou, J.; Wu, C.J.; Coleman, C.; Ng, A.; Mattson, P.; Reddi, V.J. DataPerf: Benchmarks for Data-Centric AI Development, 2023, [arXiv:cs.LG/2207.10062].
12. Zha, D.; Bhat, Z.P.; Lai, K.H.; Yang, F.; Jiang, Z.; Zhong, S.; Hu, X. Data-centric Artificial Intelligence: A Survey, 2023, [arXiv:cs.LG/2303.10158].
13. Jakubik, J.; Vössing, M.; Kühl, N.e.a. Data-Centric Artificial Intelligence. *Business & Information Systems Engineering* **2024**. doi:10.1007/s12599-024-00857-8.
14. Singh, P. Systematic review of data-centric approaches in artificial intelligence and machine learning. *Data Science and Management* **2023**, *6*, 144–157. doi:https://doi.org/10.1016/j.dsm.2023.06.001.
15. Iosifidis, V.; Ntoutsi, E. Dealing with Bias via Data Augmentation in Supervised Learning Scenarios. 2018.
16. Sharma, S.; Zhang, Y.; Ríos Aliaga, J.M.; Bouneffouf, D.; Muthusamy, V.; Varshney, K.R. Data augmentation for discrimination prevention and bias disambiguation. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 358–364.
17. Kumar, T.; Mileo, A.; Brennan, R.; Bendechache, M. Image Data Augmentation Approaches: A Comprehensive Survey and Future directions, 2023, [arXiv:cs.CV/2301.02830].

18. Alkhawaldeh, I.M.; Albalkhi, I.; Naswhan, A.J. Challenges and Limitations of Synthetic Minority Oversampling Techniques in Machine Learning. *World Journal of Methodology* **2023**, *13*, 373–378. doi:10.5662/wjm.v13.i5.373.
19. Mikołajczyk-Bareła, A. Data augmentation and explainability for bias discovery and mitigation in deep learning, 2023, [arXiv:cs.LG/2308.09464].
20. Balestriero, R.; Bottou, L.; LeCun, Y. The Effects of Regularization and Data Augmentation are Class Dependent, 2022, [arXiv:cs.LG/2204.03632].
21. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* **2002**, *16*, 321–357. doi:10.1613/jair.953.
22. Elnahas, M.; Hussein, M.; Keshk, A. Imbalanced Data Oversampling Technique Based on Convex Combination Method. *IJCI. International Journal of Computers and Information* **2022**, *9*, 15–28. doi:10.21608/ijci.2021.72508.1047.
23. Celis, L.E.; Keswani, V.; Vishnoi, N. Data preprocessing to mitigate bias: A maximum entropy based approach. Proceedings of the 37th International Conference on Machine Learning; III, H.D.; Singh, A., Eds. PMLR, 2020, Vol. 119, *Proceedings of Machine Learning Research*, pp. 1349–1359.
24. Temraz, M.; Keane, M.T. Solving the Class Imbalance Problem Using a Counterfactual Method for Data Augmentation, 2021, [arXiv:cs.LG/2111.03516].
25. He, H.; Bai, Y.; Garcia, E.; Li, S. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. 2008, pp. 1322 – 1328. doi:10.1109/IJCNN.2008.4633969.
26. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2; MIT Press: Cambridge, MA, USA, 2014; NIPS'14, p. 2672–2680.
27. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014, [http://arxiv.org/abs/1312.6114v10].
28. Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. Modeling Tabular data using Conditional GAN. Advances in Neural Information Processing Systems, 2019.
29. Patki, N.; Wedge, R.; Veeramachaneni, K. The Synthetic data vault. IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2016, pp. 399–410. doi:10.1109/DSAA.2016.49.
30. Tang, Z.; Gao, Y.; Karlinsky, L.; Sattigeri, P.; Feris, R.; Metaxas, D. OnlineAugment: Online Data Augmentation with Less Domain Knowledge. Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII; Springer-Verlag: Berlin, Heidelberg, 2020; p. 313–329. doi:10.1007/978-3-030-58571-6_19.
31. Mumuni, A.; Mumuni, F. Data augmentation: A comprehensive survey of modern approaches. *Array* **2022**, *16*, 100258. doi:https://doi.org/10.1016/j.array.2022.100258.
32. Fails, J.A.; Olsen, D.R. Interactive Machine Learning. Proceedings of the 8th International Conference on Intelligent User Interfaces; Association for Computing Machinery: New York, NY, USA, 2003; IUI '03, p. 39–45. doi:10.1145/604045.604056.
33. Kulesza, T.; Stumpf, S.; Burnett, M.; Wong, W.K.; Riche, Y.; Moore, T.; Oberst, I.; Shinsel, A.; McIntosh, K. Explanatory Debugging: Supporting End-User Debugging of Machine-Learned Programs. 2010 IEEE Symposium on Visual Languages and Human-Centric Computing; IEEE: Leganes, Madrid, Spain, 2010; pp. 41–48. doi:10.1109/VLHCC.2010.15.
34. Kulesza, T.; Burnett, M.; Wong, W.K.; Stumpf, S. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. Proceedings of the 20th International Conference on Intelligent User Interfaces; ACM: Atlanta Georgia USA, 2015; pp. 126–137. doi:10.1145/2678025.2701399.
35. Teso, S.; Alkan, O.; Stammer, W.; Daly, E. Leveraging Explanations in Interactive Machine Learning: An Overview, 2022. arXiv:2207.14526 [cs].
36. Teso, S.; Kersting, K. Explanatory Interactive Machine Learning. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society; Association for Computing Machinery: New York, NY, USA, 2019; AIES '19, p. 239–245. doi:10.1145/3306618.3314293.
37. Bhattacharya, A.; Stumpf, S.; Gosak, L.; Stiglic, G.; Verbert, K. EXMOS: Explanatory Model Steering Through Multifaceted Explanations and Data Configurations. Proceedings of the CHI Conference on Human

Factors in Computing Systems; Association for Computing Machinery: New York, NY, USA, 2024; CHI '24. doi:10.1145/3613904.3642106.

38. Muralidhar, N.; Islam, M.R.; Marwah, M.; Karpatne, A.; Ramakrishnan, N. Incorporating prior domain knowledge into deep neural networks. 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018, pp. 36–45.
39. Spinner, T.; Schlegel, U.; Schäfer, H.; El-Assady, M. explAIner: A visual analytics framework for interactive and explainable machine learning. *IEEE trans. on visualization and computer graphics* **2019**, *26*, 1064–1074.
40. Stumpf, S.; Rajaram, V.; Li, L.; Wong, W.K.; Burnett, M.; Dietterich, T.; Sullivan, E.; Herlocker, J. Interacting meaningfully with machine learning systems: Three experiments. *Int. Journal of Human-Computer Studies* **2009**, *67*, 639–662.
41. Guo, L.; Daly, E.M.; Alkan, O.K.; Mattetti, M.; Cornec, O.; Knijnenburg, B.P. Building Trust in Interactive Machine Learning via User Contributed Interpretable Rules. *27th International Conference on Intelligent User Interfaces* **2022**.
42. Honeycutt, D.R.; Nourani, M.; Ragan, E.D. Soliciting Human-in-the-Loop User Feedback for Interactive Machine Learning Reduces User Trust and Impressions of Model Accuracy, 2020, [arXiv:cs.HC/2008.12735].
43. Bhattacharya, A.; Stumpf, S.; Gosak, L.; Stiglic, G.; Verbert, K. Lessons Learned from EXMOS User Studies: A Technical Report Summarizing Key Takeaways from User Studies Conducted to Evaluate The EXMOS Platform, 2023, [arXiv:cs.LG/2310.02063].
44. Schramowski, P.; Stammer, W.; Teso, S.; Brugger, A.; Herbert, F.; Shao, X.; Luigs, H.G.; Mahlein, A.K.; Kersting, K. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence* **2020**, *2*, 476–486. doi:10.1038/s42256-020-0212-3.
45. Feuerriegel, S.; Dolata, M.; Schwabe, G. Fair AI. *Business & information systems engineering* **2020**, *62*, 379–384.
46. Bhattacharya, A.; Ooge, J.; Stiglic, G.; Verbert, K. Directive Explanations for Monitoring the Risk of Diabetes Onset: Introducing Directive Data-Centric Explanations and Combinations to Support What-If Explorations. Proceedings of the 28th International Conference on Intelligent User Interfaces; Association for Computing Machinery: New York, NY, USA, 2023; IUI '23, p. 204–219. doi:10.1145/3581641.3584075.
47. Lakkaraju, H.; Slack, D.; Chen, Y.; Tan, C.; Singh, S. Rethinking Explainability as a Dialogue: A Practitioner's Perspective, 2022, [arXiv:cs.LG/2202.01875].
48. Anik, A.I.; Bunt, A. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; ACM: Yokohama Japan, 2021; pp. 1–13. doi:10.1145/3411764.3445736.
49. Bove, C.; Aigrain, J.; Lesot, M.J.; Tijus, C.; Detyniecki, M. Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users. 27th International Conference on Intelligent User Interfaces; ACM: Helsinki Finland, 2022; pp. 807–819. doi:10.1145/3490099.3511139.
50. Lim, B.Y.; Dey, A.K.; Avrahami, D. Why and why not explanations improve the intelligibility of context-aware intelligent systems. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; Association for Computing Machinery: New York, NY, USA, 2009; CHI '09, p. 2119–2128. doi:10.1145/1518701.1519023.
51. Wang, D.; Yang, Q.; Abdul, A.; Lim, B.Y. Designing Theory-Driven User-Centric Explainable AI. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; ACM: Glasgow Scotland Uk, 2019; pp. 1–15. doi:10.1145/3290605.3300831.
52. Bhattacharya, A. Towards Directive Explanations: Crafting Explainable AI Systems for Actionable Human-AI Interactions. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24); ACM: New York, NY, USA, 2024; CHI EA '24, p. 6. doi:10.1145/3613905.3638177.