Comprehending and Mitigating Feature Bias in Machine Learning Models for Ethical AI

Youssef Hani Abdelfattah Mohamed
Department of Computer Science, Minia University, Minya, Egypt

Abstract

The critical importance of understanding and rectifying feature bias in machine learning (ML) models is pivotal in the development of fair and reliable artificial intelligence (AI) systems. This study delves into the nature and origins of feature bias, which arises when ML models base decisions on skewed or unrepresentative data, leading to potentially biased or erroneous outcomes for certain demographics. Key factors contributing to this bias include prejudices in data collection, historical and societal biases in the data, and biases inherent in the data labeling process. The ramifications of such biases are profound, potentially resulting in discriminatory practices and stereotyping, thereby diminishing the model's effectiveness in diverse real-world applications. The research presents methodologies for addressing feature bias, emphasizing the importance of diverse data sets and regular bias auditing using statistical methods to identify and quantify biases. In the realm of model development, the focus is on algorithmic fairness, including the implementation of fairness constraints or objectives during the model training process, and the careful selection and engineering of features to avoid proxies for sensitive attributes like race or gender. The paper also highlights the significance of diverse testing scenarios, independent review of model predictions, continuous monitoring through feedback loops, and regular model updates to reflect changing societal norms and values.

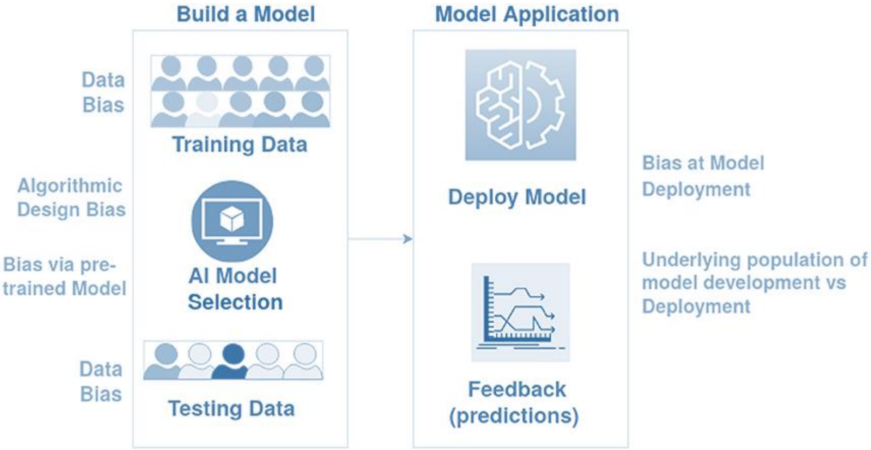
Keywords: Feature Bias, Machine Learning, Algorithmic Fairness, Ethical Artificial Intelligence, Data Diversity

Introduction

Artificial Intelligence (AI) has emerged as a transformative force in today's world, impacting various aspects of our lives, from healthcare and finance to education and transportation [1], [2]. However, the rapid proliferation of AI technologies has also raised concerns about the presence of bias within these systems. Bias in AI refers to the unfair and unjustified discrimination or favoritism that can result from the data, algorithms, or decision-making processes used in AI systems. The importance of fair and reliable AI systems cannot be overstated, as they have far-reaching implications for society, ethics, and the very foundations of our technological progress.

To understand the significance of addressing bias in AI, it is essential to first recognize that AI systems learn from the data they are trained on. This learning process is heavily dependent on the quality and diversity of the training data. If the training data is biased or unrepresentative of the real world, AI systems can inadvertently perpetuate and even amplify those biases in their decisions and predictions. For instance, if a facial recognition system is trained primarily on data that overrepresents certain racial or gender groups, it is likely to perform poorly for underrepresented groups, leading to unequal treatment and potential harm.

Figure 1. Bias at various stages of AI development



Bias in AI can manifest in various forms, including racial bias, gender bias, socioeconomic bias, and more. These biases can lead to discriminatory outcomes in fields like hiring, lending, criminal justice, and healthcare. For example, an AI-driven hiring tool that favors applicants from a particular demographic could reinforce existing disparities in the job market. Similarly, an AI algorithm used in criminal sentencing may exhibit racial bias, resulting in unjust penalties for certain racial groups. Such biases have profound consequences for individuals and communities and undermine the principles of fairness, justice, and equality.

The development and deployment of AI models are complex processes, encompassing several critical stages, each with its own potential for introducing biases. These biases can significantly impact the model's effectiveness and fairness.

The first stage is problem definition and data collection. Here, the objective of the AI model is identified, and relevant data are gathered. The way a problem is defined can greatly influence the model's focus and approach. For instance, if an AI system is designed to assess job applications, defining the criteria for a 'suitable' candidate in a narrow or biased manner could lead to a model that inadvertently discriminates against certain groups. The data collection phase is also fraught with potential biases. Data that does not comprehensively represent the target population can skew the model's understanding and predictions. An example of this is seen in facial recognition systems; if such a system is trained predominantly on images of individuals from one ethnic group, it may fail to accurately recognize faces from other ethnicities. This lack of diversity in training data is a common source of bias in AI systems [3], [4].

In the data preprocessing and feature selection stage, the collected data is cleaned and prepared for analysis, and specific features are chosen to be included in the model. This stage is crucial because decisions made here directly affect the model's learning process. Biases can be introduced through the selection of features that inadvertently encode prejudicial assumptions. For instance, using zip codes as a feature in a loan approval AI model could unintentionally introduce socio-economic biases, as zip codes can correlate with income levels and ethnic demographics [5].

The model training stage is where the AI 'learns' from the data. Here, the model is exposed to the data, and it develops algorithms to make predictions or decisions based on that data. Biases in the training data can be amplified during this phase. For example, if an AI model for credit scoring is trained on historical data where certain groups were unfairly disadvantaged, the model may perpetuate this bias, denying credit to individuals from these groups at a higher rate. It's essential at this stage to apply techniques that identify and mitigate these biases, ensuring the model's outputs are fair and unbiased.

Once the model is developed, it undergoes testing and validation. This is where the model's performance is evaluated, often using a separate dataset. Bias can manifest in this stage if the testing data is not representative of the real-world scenario in which the AI will operate. For instance, a medical diagnostic AI tested primarily on data from one country may not perform accurately when used in a different country with a different demographic profile. This can lead to misdiagnoses or missed diagnoses, particularly for underrepresented groups in the testing data.

The final stage is deployment and monitoring. Here, the AI model is put into actual use, and its performance is continuously monitored. Even after careful development and testing, biases can emerge when the model interacts with real-world data and scenarios. Continuous monitoring is crucial to identify and correct these biases. For example, an AI system used for moderating online content might initially function well, but as it encounters new types of content and user interactions, biased patterns of moderation could emerge, necessitating adjustments to the model [6], [7].

One of the main reasons bias creeps into AI systems is the bias present in the data used for training. Historical and societal biases are often reflected in data, and if not carefully curated and cleansed, these biases become ingrained in the AI algorithms. Additionally, the algorithms themselves may introduce bias, especially when they are designed without thorough consideration of potential sources of bias or with limited diversity in the development teams. To mitigate bias in AI, it is crucial to adopt a holistic approach that addresses bias at every stage of the AI system's lifecycle, from data collection and preprocessing to algorithm design and post-deployment monitoring [8].

The need for fair and reliable AI systems becomes even more critical when we consider the real-world impact of AI technologies. For instance, in the healthcare sector, AI is increasingly used for diagnosing diseases, predicting patient outcomes, and making treatment recommendations. Biased AI systems can lead to misdiagnoses or incorrect treatment plans, posing a direct risk to patients' health and well-being. Moreover, the deployment of biased AI systems in critical domains such as autonomous vehicles or aviation can have disastrous consequences, including accidents and loss of human lives.

Beyond the immediate risks, bias in AI can erode public trust in technology and exacerbate societal divisions. When individuals perceive AI systems as unfair or discriminatory, they may resist their adoption or demand increased regulation, slowing down innovation and progress. Additionally, bias in AI can reinforce stereotypes and perpetuate systemic inequalities, making it even harder to address long-standing societal issues. To build trust and ensure the widespread acceptance of AI, it is imperative to prioritize fairness and reliability in AI development.

The quest for fair and reliable AI systems involves multiple dimensions. Firstly, it requires diverse and representative data. Data collection efforts should encompass a wide range of demographic groups, ensuring that minority populations and underrepresented communities are not overlooked. Furthermore, data should be carefully curated to eliminate biased or unrepresentative samples. This process often involves data preprocessing techniques like data augmentation and oversampling to balance the dataset and reduce bias [9], [10].

Secondly, AI algorithms must be designed with fairness in mind. Developers should consider various fairness metrics and incorporate them into the evaluation of their models. These metrics can help identify and quantify bias in AI systems, enabling developers to make informed adjustments. Techniques like adversarial training and reweighting of samples can also be used to reduce bias in AI algorithms. However, it is crucial to strike a balance between fairness and other performance metrics, as overly aggressive fairness constraints can lead to reduced accuracy and utility [11].

Transparency and interpretability are another key aspect of building fair and reliable AI systems. To gain the trust of users and stakeholders, AI models should be explainable, allowing individuals to understand the rationale behind the system's decisions. This not only aids in identifying and rectifying bias but also ensures accountability when AI systems make errors or exhibit unfair behavior. Techniques like model interpretability, feature importance analysis, and algorithmic transparency play a crucial role in achieving this goal.

Ongoing monitoring and evaluation of AI systems in real-world scenarios are essential to ensure their continued fairness and reliability. This involves setting up mechanisms for feedback, auditing, and continuous improvement. By regularly assessing how AI systems perform in different contexts and for different user groups, developers can identify and rectify bias that may emerge over time or due to changing circumstances. Such iterative processes are fundamental to maintaining fairness and reliability in the ever-evolving landscape of AI applications.

Apart from technical considerations, fostering diversity and inclusivity in AI development teams is vital. Diverse teams are more likely to recognize and address bias, as they bring different perspectives and experiences to the table. Moreover, involving stakeholders from various backgrounds, including ethicists, social scientists, and community representatives, in the development process can help uncover potential sources of bias and ensure that AI systems align with societal values.

Ethical considerations also play a pivotal role in the development of fair and reliable AI systems. Developers and organizations should adhere to ethical guidelines and principles that prioritize human rights, fairness, and non-discrimination. This involves making conscious decisions about the goals and consequences of AI systems, acknowledging their potential for harm, and taking steps to mitigate that harm. Ethical AI frameworks like the principles outlined in the Universal Declaration on Artificial Intelligence are valuable resources in this regard.

Regulation and policy also have a role to play in promoting fairness and reliability in AI. Governments and regulatory bodies must establish clear guidelines and standards for AI development and deployment. These regulations should include requirements for transparency, fairness assessments, and accountability mechanisms. Implementing such policies can create a level playing field for AI developers and encourage the adoption of best practices across industries.

To underscore the importance of fairness and reliability in AI systems, it is worth considering the ethical and philosophical foundations that underpin these principles. AI systems, while powerful tools, are creations of human ingenuity, and as such, they should reflect the values and aspirations of society. Fairness in AI aligns with the moral principle of treating all individuals with equal dignity and respect, regardless of their race, gender, or socioeconomic background. Reliability, in turn, upholds the ethical obligation to ensure that AI systems perform their intended functions accurately and without causing harm.

Moreover, the pursuit of fairness and reliability in AI systems resonates with broader societal goals such as social justice, equity, and human rights. These principles are enshrined in international agreements and constitutions across the world. By embedding fairness and reliability in AI development, we uphold the ideals of a just and equitable society where every individual has the opportunity to thrive and reach their full potential [12].

Bias in AI is a pressing issue that has far-reaching consequences for individuals, communities, and society as a whole. Fair and reliable AI systems are essential to prevent discrimination, promote equity, and ensure that AI technologies fulfill their promise of improving human lives. Achieving fairness and reliability in AI involves a multifaceted approach, encompassing data collection, algorithm design, transparency, diversity, ethics, and regulation. It requires a collective effort from researchers, developers, policymakers, and the broader society to build a future where AI systems truly reflect the values of fairness, reliability, and respect for all. By prioritizing these principles, we can harness the full potential of AI while minimizing the risks and harms associated with bias in technology.

Comprehension of Feature Bias

Feature bias in machine learning occurs when a model's decisions are influenced by biased or unrepresentative data, which can lead to biased or inaccurate predictions for specific groups. This bias can have its roots in various aspects of the data and model training process.

One of the primary sources of feature bias is data collection. When data is collected, it may inadvertently overrepresent a particular demographic or group, leading to an imbalance in the dataset. This overrepresentation can result from various factors, including the availability of data from specific sources or populations, convenience in data collection, or historical biases in data collection practices. When a dataset is heavily skewed towards one group, it can diminish the model's effectiveness for other, underrepresented groups. This skewed representation may lead to the model making better predictions for the overrepresented group and performing poorly for others.

Historical biases also play a significant role in the emergence of feature bias. If the training data used to develop the model contains historical or societal prejudices, the model may inadvertently learn and perpetuate these biases. For example, if historical data reflects discriminatory practices, such as unequal treatment based on race or gender, the model may incorporate these biases into its decision-making process. Consequently, it may make predictions that unfairly disadvantage certain groups or reinforce existing stereotypes [13], [14].

Label bias is another factor contributing to feature bias. The way data is labeled or categorized can influence the model's training and its subsequent predictions. Biased labeling can occur when human annotators introduce their own biases or when existing labels contain inherent biases. This can skew the model's perception of the data, causing it to make biased predictions that align with the labeled categories [15], [16].

The implications of feature bias in machine learning are significant and wide-ranging. When feature bias is present in a model, it can lead to discriminatory practices in various real-world applications. For instance, biased AI systems used in hiring processes may favor one demographic group over others, perpetuating disparities in employment opportunities. In healthcare, biased predictive models may provide less accurate diagnoses or treatment recommendations for certain populations, leading to unequal healthcare outcomes. Furthermore, feature bias can reinforce harmful stereotypes, exacerbate societal inequalities, and erode trust in AI systems [17].

The recognition and mitigation of feature bias are crucial in the development of fair and equitable machine learning models. Addressing feature bias requires careful consideration of data collection practices, thorough data preprocessing to reduce biases, and the incorporation of fairness-aware algorithms and evaluation metrics into the model development process. Failure to address feature bias can have detrimental consequences, undermining the potential benefits of AI and perpetuating injustices in our increasingly automated world.

Rectification of Feature Bias

Data Analysis plays a pivotal role in ensuring the integrity and fairness of machine learning models. One of the fundamental aspects of data analysis is the utilization of Diverse Data Sets. It is imperative to ensure that the data used for training encompasses a broad spectrum of demographics and scenarios. By including diverse data, we reduce the risk of creating models that are biased or discriminatory. For instance, when developing a predictive model for loan approvals, using data from various income levels, ethnicities, and educational backgrounds ensures that the model is less likely to favor one group over another. Diverse data sets help in creating models that are fair and equitable, reflecting the real-world diversity of the populations they serve.

Bias Auditing is another critical component of data analysis. Regular audits of the data are essential to detect potential biases that may exist in the dataset. Statistical techniques can be employed to identify and measure these biases. Bias auditing allows data scientists and researchers to uncover disparities and imbalances in the data, whether they stem from historical biases in data collection or other sources. Identifying bias at this stage is crucial, as it provides an opportunity to mitigate it during the model development process. Without proper bias auditing, machine learning models may inadvertently perpetuate unfair or discriminatory outcomes, leading to real-world harm and injustice [18], [19].

Moving from data analysis to Model Development, Algorithmic Fairness takes center stage. The implementation of algorithms aimed at minimizing bias is crucial for building fair and ethical AI systems. This may encompass the integration of fairness constraints or objectives within the model training regimen. For example, fairness-aware algorithms can be designed to ensure that the predictions made by the model are equitable across different demographic groups. These algorithms strive to strike a balance between accuracy and fairness, ensuring that no particular group is systematically disadvantaged. Algorithmic fairness is a proactive approach to addressing bias and discrimination, and it is an essential consideration in the development of responsible AI systems [20].

Feature Selection is another critical aspect of Model Development that plays a significant role in preventing bias. When selecting and engineering features for the model, caution must be exercised to avoid including attributes that could serve as surrogates for sensitive attributes, such as race or gender. These sensitive attributes are often referred to as "protected attributes," and their direct inclusion in the model can lead to biased outcomes. Feature selection methods should be used to identify and exclude such attributes or to carefully design features that mitigate the risk of bias. By considering feature selection through the lens of fairness, developers can create models that are less likely to discriminate based on sensitive characteristics, thus promoting fairness and equity in AI systems.

Data analysis and model development are integral parts of building fair and ethical machine learning models. Diverse Data Sets and Bias Auditing in the data analysis phase help in ensuring that the data used for training is representative and free from bias. In Model Development, Algorithmic Fairness techniques are employed to minimize bias and promote equitable outcomes. Additionally, Feature Selection is crucial to prevent the inclusion of attributes that may inadvertently introduce bias into the model. By following these principles and integrating fairness considerations into every stage of the machine learning pipeline, we can work towards creating AI systems that are fair, ethical, and equitable, serving the diverse needs of society while minimizing the risk of harm and discrimination.

Testing and Validation are crucial phases in the development of machine learning models to ensure their fairness and reliability. To begin with, Diverse Testing Scenarios are essential. It is vital to test the model across a diverse array of scenarios to assure its equitable performance. Models should be evaluated on a wide range of inputs and conditions to ensure that they do not favor any particular group or demographic. For instance, a facial recognition system should be tested with faces from different racial backgrounds and under various lighting conditions to ensure that it performs fairly for all users. Diverse testing scenarios help in identifying potential biases and shortcomings in the model's behavior [21].

Independent Review is another important aspect of testing and validation. The predictions made by the model should be subject to evaluation by a varied group of independent experts who can assess the model's performance and detect potential biases. These experts can provide valuable insights and objective assessments, helping to uncover any hidden biases or unfair outcomes. Independent review adds an extra layer of scrutiny and accountability, ensuring that AI systems are held to high standards of fairness and reliability.

Moving beyond the testing phase, Continuous Monitoring is essential to maintain the fairness and reliability of AI systems over time. This involves the establishment of Feedback Loops, where the model's performance is consistently monitored, and feedback is collected from real-world usage. Feedback mechanisms allow developers to learn from the model's performance in various contexts and make necessary adjustments to address biases or improve fairness. Continuous monitoring ensures that AI systems remain aligned with their intended goals and adapt to changing conditions.

Adaptation is another critical aspect of continuous monitoring. AI models should be regularly updated to align with evolving societal norms and values. What may be considered fair and unbiased today may not hold true in the future. By adapting and updating models, developers can ensure that AI systems stay relevant and continue to serve the best interests of society. For example, as societal understanding of fairness evolves, models used in decision-making should be updated to reflect these changes and avoid perpetuating outdated biases.

In conclusion, testing, validation, and continuous monitoring are integral to the development of fair and reliable machine learning models. Diverse Testing Scenarios and Independent Review help identify and rectify biases and ensure that models perform equitably across various scenarios and demographics. Continuous Monitoring through

Feedback Loops and Adaptation ensures that AI systems remain fair and adaptable to changing societal norms and values. By following these principles and incorporating them into the lifecycle of AI systems, we can build and maintain AI technologies that are not only technically proficient but also ethically responsible, serving the diverse needs of society while minimizing the risk of bias and discrimination [22].

Ethical and Legal Considerations play a critical role in the development and deployment of machine learning models, particularly when it comes to ensuring fairness, accountability, and responsible AI [23].

Firstly, Compliance with legal standards is of utmost importance. Developers and organizations must ensure that the model adheres to all relevant legal regulations concerning discrimination and privacy. This includes laws such as the Fair Housing Act, the Equal Credit Opportunity Act, and data protection regulations like the General Data Protection Regulation (GDPR). Compliance not only safeguards against legal repercussions but also upholds the principles of fairness and non-discrimination in AI systems. Ensuring that the model operates within the bounds of the law is a fundamental ethical consideration.

Transparency is another essential ethical consideration, especially in sensitive applications of AI. Maintaining transparency regarding the decision-making processes of the model is crucial. Users and stakeholders should have a clear understanding of how the model arrives at its predictions or decisions. In cases where AI systems are used for critical decisions like healthcare or criminal justice, transparency is vital for accountability and trust. Transparent AI systems allow for scrutiny, which can help identify and address biases or discriminatory practices. By providing explanations and insights into its operations, AI can be more ethically responsible and aligned with societal values.

Ethical and legal considerations are integral to the development of machine learning models that are fair, accountable, and responsible. Compliance with legal standards ensures that models operate within the boundaries of the law, promoting fairness and non-discrimination. Transparency, especially in sensitive applications, fosters accountability and trust by allowing users and stakeholders to understand the model's decision-making processes. These considerations are essential to the responsible use of AI and the preservation of ethical principles in technology development.

Conclusion

Comprehension of Feature Bias involves understanding the subtle yet significant ways in which machine learning models can be influenced by biased or unrepresentative data. This phenomenon, known as feature bias, can lead to prejudiced or erroneous predictions, especially for certain groups. It's a multifaceted issue that spans across various stages of machine learning development, from data collection to model application, and requires careful consideration to mitigate its effects [24].

The origins of feature bias are diverse. During data collection, biases can creep in if the dataset disproportionately represents certain demographics, thus diminishing the model's accuracy or fairness for other groups. Historical bias is another source, where the model inadvertently learns and perpetuates societal or historical prejudices that are embedded in the data. Additionally, the way data is labeled or categorized can introduce bias, affecting how the model interprets and responds to different features [25], [26].

The implications of feature bias are far-reaching. It can lead to discriminatory practices, reinforcing stereotypes and hindering the model's effectiveness in diverse real-world scenarios. This is particularly concerning in areas like healthcare, finance, and law enforcement, where biased predictions can have serious, life-altering consequences.

To rectify feature bias, a multi-pronged approach is necessary. Starting with data analysis, it's crucial to ensure the data set is diverse and representative of various demographics and scenarios. Regular bias auditing using statistical techniques helps in identifying and measuring potential biases in the data. In the model development phase, algorithms should be designed to minimize bias. This might include integrating fairness constraints or objectives into the training process. Careful feature selection is also important, avoiding features that might act as proxies for sensitive attributes like race or gender.

Testing and validation are key in ensuring a model's unbiased performance. The model should be tested across a wide range of scenarios to evaluate its fairness and effectiveness. Independent review by a diverse group of experts can help in identifying any overlooked biases. Additionally, continuous monitoring of the model is essential. Establishing feedback loops allows for ongoing learning and adjustment based on the model's performance. The model should also be regularly updated to keep pace with changing societal norms and values.

Finally, ethical and legal considerations are paramount. Ensuring the model complies with legal standards related to discrimination and privacy is crucial. Transparency in the model's decision-making processes, especially in sensitive applications, is necessary to maintain public trust and accountability.

In conclusion, comprehending and addressing feature bias in machine learning is a complex but essential task. It requires a thorough and continuous effort across all stages of model development and application, ensuring that these powerful tools are used ethically and effectively for the benefit of all.

References

[1] D. Mahapatra, A. Poellinger, and M. Reyes, “Interpretability-Guided Inductive Bias For Deep Learning Based Medical Image,” *Med. Image Anal.*, vol. 81, p. 102551, Oct. 2022.

[2] A. Goyal and Y. Bengio, “Inductive biases for deep learning of higher-level cognition,” *Proceedings of the Royal Society A*, 2022.

[3] N. Sourlos, J. Wang, Y. Nagaraj, P. van Ooijen, and R. Vliegenthart, “Possible Bias in Supervised Deep Learning Algorithms for CT Lung Nodule Detection and Classification,” *Cancers* , vol. 14, no. 16, Aug. 2022.

[4] T. Feldman and A. Peake, “End-To-End Bias Mitigation: Removing Gender Bias in Deep Learning,” *arXiv [cs.LG]*, 06-Apr-2021.

[5] F. N. U. Jirigesi, “Personalized Web Services Interface Design Using Interactive Computational Search.” 2017.

[6] Y. Liu, J. Wang, J. Li, H. Song, and T. Yang, “Zero-bias deep learning for accurate identification of Internet-of-Things (IoT) devices,” *IEEE Internet of*, 2020.

[7] I. Shrier and R. W. Platt, “Reducing bias through directed acyclic graphs,” *BMC Med. Res. Methodol.*, vol. 8, p. 70, Oct. 2008.

[8] F. Jirigesi, A. Truelove, and F. Yazdani, “Code Clone Detection Using Representation Learning,” 2019.

[9] S. Kelley, A. Ovchinnikov, D. R. Hardoon, and A. Heinrich, “Antidiscrimination Laws, Artificial Intelligence, and Gender Bias: A Case Study in Nonmortgage Fintech Lending,” *M&SOM*, vol. 24, no. 6, pp. 3039–3059, Nov. 2022.

[10] Y. McQuinlan *et al.*, “An investigation into the risk of population bias in deep learning autocontouring,” *Radiother. Oncol.*, vol. 186, p. 109747, Sep. 2023.

[11] J. Gesi, J. Li, and I. Ahmed, “An empirical examination of the impact of bias on just-in-time defect prediction,” in *Proceedings of the 15th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 2021, pp. 1–12.

[12] A. Groce *et al.*, “Evaluating and improving static analysis tools via differential mutation analysis,” in *2021 IEEE 21st International Conference on Software Quality, Reliability and Security (QRS)*, 2021, pp. 207–218.

[13] S. Vaid, R. Kalantar, and M. Bhandari, “Deep learning COVID-19 detection bias: accuracy through artificial intelligence,” *Int. Orthop.*, vol. 44, no. 8, pp. 1539–1542, Aug. 2020.

[14] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars, “A Deeper Look at Dataset Bias,” in *Domain Adaptation in Computer Vision Applications*, G. Csurka, Ed. Cham: Springer International Publishing, 2017, pp. 37–55.

[15] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2019.

[16] K. Gade, S. C. Geyik, K. Kenthapadi, V. Mithal, and A. Taly, “Explainable AI in Industry,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, AK, USA, 2019, pp. 3203–3204.

[17] J. Gesi *et al.*, “Code smells in machine learning systems,” *arXiv preprint arXiv:2203.00803*, 2022.

[18] M. Nagendran *et al.*, “Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies,” *BMJ*, vol. 368, p. m689, Mar. 2020.

[19] D. Pedreschi, F. Giannotti, R. Guidotti, A. Monreale, S. Ruggieri, and F. Turini, “Meaningful Explanations of Black Box AI Decision Systems,” *AAAI*, vol. 33, no. 01, pp. 9780–9784, Jul. 2019.

[20] J. Gesi, X. Shen, Y. Geng, Q. Chen, and I. Ahmed, “Leveraging Feature Bias for Scalable Misprediction Explanation of Machine Learning Models,” in *Proceedings of the 45th International Conference on Software Engineering (ICSE)*, 2023.

[21] J. Gesi, H. Wang, B. Wang, A. Truelove, J. Park, and I. Ahmed, “Out of Time: A Case Study of Using Team and Modification Representation Learning for Improving Bug Report Resolution Time Prediction in Ebay,” *Available at SSRN 4571372*.

[22] P. A. Noseworthy *et al.*, “Assessing and Mitigating Bias in Medical Artificial Intelligence: The Effects of Race and Ethnicity on a Deep Learning Model for ECG Analysis,” *Circ. Arrhythm. Electrophysiol.*, vol. 13, no. 3, p. e007988, Mar. 2020.

[23] Y. Huang *et al.*, “Behavior-driven query similarity prediction based on pre-trained language models for e-commerce search,” 2023.

[24] S. Gerke, T. Minssen, and G. Cohen, “Chapter 12 - Ethical and legal challenges of artificial intelligence-driven healthcare,” in *Artificial Intelligence in*

Healthcare, A. Bohr and K. Memarzadeh, Eds. Academic Press, 2020, pp. 295–336.

[25] N. Naik *et al.*, “Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility?,” *Front Surg*, vol. 9, p. 862322, Mar. 2022.

[26] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, “Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 117, no. 23, pp. 12592–12594, Jun. 2020.