Fairness in Machine Learning Meets with Equity in Healthcare

Shaina Raza[1]*, **Parisa Osivand Pour**[1], **Syed Raza Bashir**[2]

[1] Vector Institute for Artificial Intelligence, Toronto, ON, Canada
[2] Toronto Metropolitan University, Toronto, ON, Canada
{shaina.raza, parisa.osivand}@vectorinstitute.ai, syedraza.bashir@torontomu.ca

Abstract

With the growing utilization of machine learning in healthcare, there is increasing potential to enhance healthcare outcomes. However, this also brings the risk of perpetuating biases in data and model design that can harm certain demographic groups based on factors such as age, gender, and race. This study proposes an artificial intelligence framework, grounded in software engineering principles, for identifying and mitigating biases in data and models while ensuring fairness in healthcare settings. A case study is presented to demonstrate how systematic biases in data can lead to amplified biases in model predictions, and machine learning methods are suggested to prevent such biases. Future research aims to test and validate the proposed ML framework in real-world clinical settings to evaluate its impact on promoting health equity.

Introduction

Machine learning (ML) offers immense potential to significantly enhance patient outcomes and transform the landscape of clinical healthcare (Thomasian, Eickhoff, and Adashi 2021). Utilizing its analytical and predictive capabilities, ML can help reveal disease patterns and trends, and optimize patient care. However, it is important to proceed with caution when leveraging ML in healthcare. This is because inherent biases and inequalities in the data may result in discrimination, which could result in worsening of preexisting health disparities (McNeely, Schintler, and Stabile 2020). For example, a model trained on biased data might inaccurately predict a higher risk of heart disease for specific racial or ethnic groups, leading to unequal treatment opportunities and poorer health outcomes (Obermeyer et al. 2019).

In the context of ML, the term "bias" refers to skewed outcomes caused by errors in the modeling process (Fletcher, Nakeshimana, and Olubeko 2021) . This often occurs when training data is unrepresentative or contains systemic errors, leading the model to learn and potentially replicate these biases in its predictions. "Disparity" in healthcare indicates inequalities in health status, healthcare access, or healthcare quality across different groups (Raza 2022). It is very important to minimize these biases and disparities when applying ML to healthcare, to ensure equitable outcomes.

Health equity (Sikstrom et al. 2022) is a core principle in clinical healthcare that seeks to eliminate differences in health outcomes and access to equal healthcare among various populations. This principle aims to ensure that all individuals, regardless of their demographic or socio-economic background, have equal opportunities to access care and maintain or improve their health. Both the World Health Organization (WHO) and the United Nations (UN) prioritize health equity as a critical element of their missions to enhance global health outcomes. This motivates us to pursue research in this domain.

In this study, we introduce an Artificial Intelligence (AI) framework designed to ensure that ML models produce unbiased and equitable predictions for all populations. Specifically, we integrate software engineering principles into the framework to improve its modularity, maintainability, and scalability, making it adaptable and efficient for various applications. Our goal is to introduce fairness in the healthcare setting through ML. The term fairness typically refers to the algorithm's ability to make decisions and predictions without unjust bias or discrimination (Rajkomar et al. 2018). Fairness is a critical aspect of responsible AI and ML practices, especially in sensitive areas like healthcare.

We put forth a fair ML framework, in this work, that is rooted in software engineering principles. Following that, we present a healthcare case study that demonstrates how biases can exacerbate disparities in healthcare access and outcomes. We also detail how our suggested framework can aid in advancing health equity. It is our hope that integrating this framework into healthcare systems will promote equal health opportunities and outcomes.

Previous Works

In the realm of healthcare, researchers have explored the potential of AI and its capacity for ensuring fairness and equity. Rajkomar et al. (Rajkomar et al. 2018) highlighted the importance of fairness in clinical care and introduced research guidelines and technical solutions to combat biases through ML . Fletcher et al. (Fletcher, Nakeshimana, and Olubeko 2021) conducted research on the global health context, particularly in Low- and Middle-Income Countries

*Corresponding author.

(LMICs), proposing three criteria—appropriateness, fairness, and bias—to evaluate ML for healthcare. Raza (Raza 2022) presented a review on the challenges for ML within a general view of public health and its influences. Thomasian et al. (Thomasian, Eickhoff, and Adashi 2021) urged for policy-level consensus on algorithmic bias and providing principles for mitigating bias in healthcare. Wesson et al. (Wesson et al. 2022) looked the potential benefits and drawbacks of using big data in research, emphasizing the importance of an equity lens in health.

Sikstrom et al. (Sikstrom et al. 2022) conducted literature survey on fairness in AI and ML, striving to operationalize fairness in medicine. Concurrently, Gervasi et al. (Gervasi et al. 2022) explored fairness, equity, and bias in ML algorithms within the health insurance industry. Obermeyer et al. (Obermeyer et al. 2019) uncovered racial bias in a commercial algorithm used for identifying high-risk patients, emphasizing the need to address racial bias in ML pipelines. The AI Now Institute (Now 2021) delved into the social implications of AI and ML, publishing works on fairness, accountability, and transparency, including healthcare ML pipelines. Google AI for Social Good program (AI 2023) are also developing tools and resources like the What-If Tool and Fairness Indicators to assist practitioners in identifying and mitigating biases in ML pipelines (Huang et al. 2021; Raza, Reji, and Ding 2022; Sikstrom et al. 2022). These works high-light the growth of ML in healthcare. Nevertheless, there is need for engagement with fairness, bias, and ML processes in healthcare.

Our work differentiates from previous research by offering a practical, end-to-end framework for implementing fair ML in healthcare. Our work is rooted in software engineering principles and focuses on specific fairness considerations in healthcare. Different from much of the existing work that mainly focuses on theoretical guidelines or principles, our research offers a tool that can have real-world applicability and continuous improvement.

Proposed Framework

We propose an AI framework, shown in Figure 1 and the steps given in Algorithm 1, that integrates software engineering principles with fairness in ML. The goal of this work is to enhance modularity, maintainability, and scalability. The steps of our proposed framework are as:

Actor Identification: Identify the key actors and their roles is one of the fundamental steps in software engineering principles. Understanding the users is crucial in this process as it provides context and direction for the subsequent steps in the framework.

Requirements Analysis: Identify the problem that we aim to solve in healthcare and determine the fairness requirements specific to the context. For example, to understand the ethical, legal, and social implications of the solution (Cordeiro 2021; Lu et al. 2022) and set goals to mitigate potential biases and promote equitable outcomes.

Data Collection: Collect diverse and representative data samples that covers various demographic (Tramer et al. 2017), to ensure the model generalizability. It is also important to ensure data privacy and security standards are met while acquiring and storing data (Raza and Schwartz 2023; Bashir et al. 2022)

Algorithm 1: Fairness in ML for healthcare

Input: Data set D, ML algorithm A
Parameter: Fairness metric F, Threshold T
Output: Trained ML model M, Fairness evaluation E

1: Pre-process data set D to get D'.
2: Split D' into training, validation, and test sets: T_{train}, T_{val}, and T_{test}.
3: Apply fairness pre-processing to get $T_{\text{train\_balanced}}$.
4: Initialize ML algorithm A with parameters P.
5: Train ML model M using $T_{\text{train\_balanced}}$.
6: Apply fairness in-processing on M during training.
7: Validate M on T_{val}, evaluate performance metrics and fairness metric F.
8: **while** $F > T$ **do**
9: Tune hyperparameters P of ML algorithm A.
10: Retrain M using $T_{\text{train\_balanced}}$ and updated P.
11: Validate M on T_{val}, re-evaluate performance metrics and fairness metric F.
12: **end while**
13: Evaluate M on T_{test} to get final performance metrics and fairness metric F.
14: Apply fairness post-processing on M if necessary.
15: Deploy M in production environment.
16: Monitor M performance and fairness metric F on new data.
17: **if** $F > T$ **then**
18: Update M and repeat steps 8-13.
19: **end if**
20: **return** M, E

Data Pre-processing: Apply best practices to clean, normalize, and transform the data. Implement fairness pre-processing techniques such as re-sampling (Drummond, Holte et al. 2003), re-weighting (Kamiran and Calders 2012), or editing feature values (Hardt, Price, and Srebro 2016) to reduce potential biases in the data set.

Feature Selection and Engineering: Identify relevant features that impact the target outcome and avoid features that might introduce biases (Ahmad, Teredesai, and Eckert 2018). Apply domain knowledge to create meaningful features that contribute to a fair model.

Model Selection and Training: Choose a suitable ML algorithm for the problem at hand, considering software engineering principles like modularity, scalability, and maintainability (Raza, Reji, and Ding 2022). Employ in-processing techniques such as fair classification (Kamiran and Calders 2012), clustering (Chierichetti et al. 2017), adversarial learning (Madras et al. 2018), and counterfactual fair learning (Kusner et al. 2017) algorithms to promote fairness during model training.

Model Validation and Evaluation: Assess the model performance using standard evaluation metrics as well as fairness-specific metrics like disparate impact (Feldman et al. 2015), demographic parity (Madras et al. 2018), or

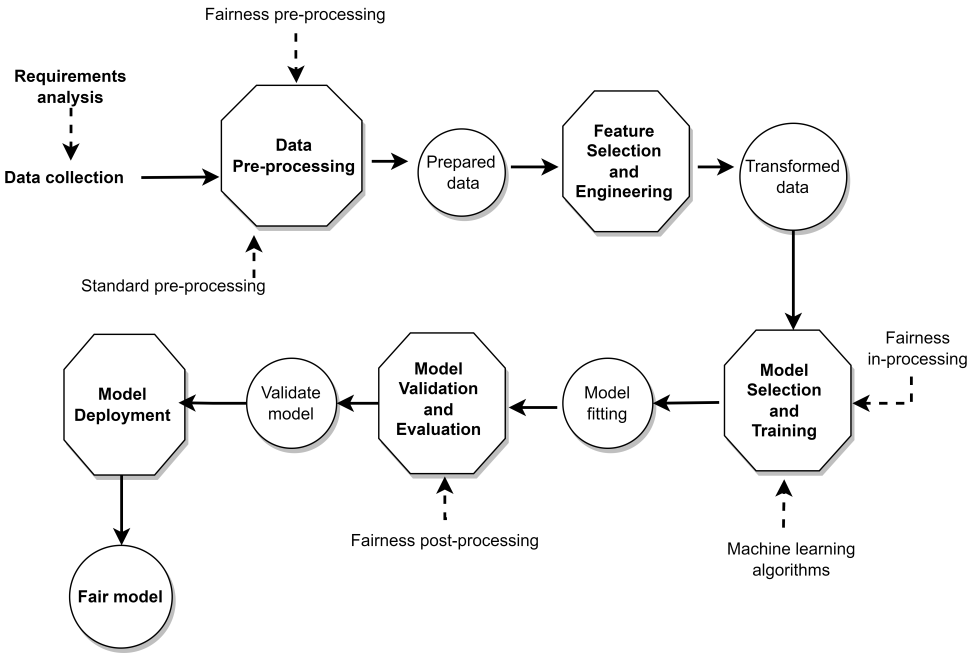

Figure 1: Proposed Framework

equalized odds (Gölz, Kahng, and Procaccia 2019). Optimize the model using hyperparameter tuning, and apply post-processing fairness techniques like counterfactual analysis (Kusner et al. 2017) or calibration (Pleiss et al. 2017) to adjust the predictions as needed.

Model Deployment and Monitoring: Deploy the model in a production environment while adhering to software engineering best practices for continuous integration, continuous deployment, and monitoring (Alanazi 2022). Regularly evaluating the model performance on new data, ensuring its fairness and generalizability over time is also important. Given that software engineering relies on empirical science for validity, understanding the users and their feedback is also important to validate our framework approach.

Case Study

Case study: We take a case study (Ting et al. 2017) of Diabetic retinopathy (DR) using ML. DR is a complication of diabetes that can lead to vision loss and blindness if not detected and treated early. The prevalence of diabetes continues to rise globally, and early detection of DR is crucial for timely intervention and preventing vision loss. Fundus photography is a widely used technique to screen for DR, but the manual examination of these images can be time-consuming and subject to inter-observer variability. ML algorithms can automate this process, making it more efficient and consistent, as shown in this work (Ting et al. 2017).

Objective: Our objective in this paper is to propose a fair and unbiased version of the original ML method (Ting et al. 2017) for early detection of diabetic retinopathy.

Data Collection and Pre-processing: A diverse and representative data set of fundus images was collected from various sources, ensuring the inclusion of different demo-graphic groups such as age, gender, and ethnicity. The data was pre-processed, including cleaning, normalization, and transformation. Fairness-enhancing preprocessing techniques: re-sampling (Huang et al. 2021) and re-weighting (Feldman et al. 2015) were applied to balance the data set and mitigate potential biases.

Feature Selection and Engineering: Domain experts identified relevant features for the detection of diabetic retinopathy, such as blood vessel structure, hemorrhages, and microaneurysms. Feature engineering techniques were applied to extract meaningful information from fundus images while avoiding features that might introduce bias.

Model Selection and Training: A convolutional neural network (CNN) (Mallat 2016) was selected as the ML algorithm, considering its effectiveness in image analysis tasks and alignment with software engineering principles like modularity and scalability. Fair classification (Dwork et al. 2012) and adversarial learning (Zhang, Lemoine, and Mitchell 2018) techniques were applied during the model training to ensure fairness and unbiased predictions.

Model Validation and Evaluation: The model was evaluated using standard metrics such as accuracy, precision, and recall, as well as fairness-specific metrics like demographic parity and equalized odds. Post-processing fairness techniques like counterfactual analysis (Kusner et al. 2017) and calibration (Pleiss et al. 2017) were applied as needed to adjust the predictions and ensure fairness across demographic groups.

Model Deployment and Monitoring: The CNN model was deployed in a production environment for continuous integration, continuous deployment, and monitoring. The model's performance and fairness were regularly evaluated

on new data, and updates were made as needed to maintain its fairness and generalizability over time.

Outcome: The fair ML-based system improved the efficiency and consistency of DR screening, reducing the workload of healthcare professionals and enabling timely intervention. By ensuring equitable predictions across diverse demographic groups (Tramer et al. 2017), the system contributed to health equity and reduced the risk of vision loss in diabetic patients (Raza 2022).

Discussion

The proposed framework and its application to the DR case study illustrate the promising potential of integrating software engineering principles with fairness in ML for healthcare. In this context, we highlight several critical observations and implications.

Achieving Fairness in Machine Learning: The integration of fairness-enhancing techniques during pre-processing, in-processing, and post-processing stages were essential to mitigate potential biases and ensure fairness in the DR detection model. These techniques worked in harmony with standard ML processes, ensuring their applicability in other healthcare contexts. However, it's worth noting that fairness is a dynamic concept, dependent on the specific healthcare problem and demographics (Dwork et al. 2012). Therefore, the selection and application of fairness techniques require domain knowledge and an understanding of the specific fairness requirements.

Role of Software Engineering Principles: Adopting software engineering principles not only enhanced the modularity, maintainability, and scalability of our ML framework but also facilitated the integration of fairness techniques. By treating the ML model as a software product, we were able to design and develop the framework more efficiently and effectively. This allowed us to monitor and adjust the model continuously, ensuring that fairness and performance were maintained over time.

Health Equity in Practice: By developing a fair and unbiased DR detection model, we demonstrated the practical application of health equity in ML for healthcare. The model made equitable predictions across diverse demographic groups, contributing to equal health opportunities and outcomes for diabetic patients. This underscores the importance of fairness in healthcare ML, not only as a theoretical concept but also as a practical tool for promoting health equity.

Collaboration Across Disciplines: This work exemplified the benefits of collaboration across disciplines. The integration of insights and methodologies from software engineering, ML, healthcare, and ethics was essential to the successful development of the proposed framework. This underscores the importance of interdisciplinary collaboration in addressing complex problems like fairness in ML for healthcare.

Despite the encouraging results from our framework and the DR case study, several challenges and limitations need to be addressed in future work. For example, it's essential to understand the legal and ethical implications of deploying such frameworks, especially regarding data privacy and informed consent (Tramer et al. 2017). Additionally, the selection and definition of fairness metrics can be subjective and may differ across contexts, requiring further investigation.

In future work, we envision refining and expanding our proposed framework to address these challenges. We also plan to apply our framework to other healthcare problems to further validate its efficacy and versatility. Furthermore, we hope to stimulate discussions and collaborations with stakeholders from various disciplines to contribute to the ongoing research and development of fair ML for healthcare. Ultimately, our goal is to contribute to health equity and improve patient outcomes using ML, while upholding the principles of fairness and justice.

Conclusion

We present an approach aimed at uncovering and addressing biases present in healthcare data, with the goal of promoting equitable solutions. The case study examined provides a foundation, suggesting that the proposed framework can effectively identify biases and apply suitable fairness methods to assess potential discrimination and generate fairer outcomes. To maximize benefits from this framework, it is crucial to prioritize fairness in all aspects of model design, deployment, and evaluation. This study has some limitations, for example, the lack of real-world empirical evidence supporting the effectiveness of the proposed AI framework. Further empirical research and real-world validation are needed to verify the proposed framework efficacy.

Acknowledgments

Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute www.vectorinstitute.ai/#partners

References

Ahmad, M. A.; Teredesai, A.; and Eckert, C. 2018. Interpretable machine learning in healthcare. In *Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018*, 447. ISBN 9781538653777.

AI, G. 2023. Google AI Social Good. https://ai.google/responsibility/social-good/. Accessed: 2023-07-28.

Alanazi, A. 2022. Using machine learning for healthcare challenges and opportunities. *Informatics in Medicine Unlocked*, 30: 100924.

Bashir, S. R.; Raza, S.; Kocaman, V.; and Qamar, U. 2022. Clinical Application of Detecting COVID-19 Risks: A Natural Language Processing Approach. *Viruses*, 14(12).

Chierichetti, F.; Kumar, R.; Lattanzi, S.; and Vassilvitskii, S. 2017. Fair clustering through fairlets. *Advances in neural information processing systems*, 30.

Cordeiro, J. V. 2021. Digital technologies and data science as health enablers: an outline of appealing promises and compelling ethical, legal, and social challenges. *Frontiers in Medicine*, 8: 647897.

Drummond, C.; Holte, R. C.; et al. 2003. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, 1–8.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268.

Fletcher, R. R.; Nakeshimana, A.; and Olubeko, O. 2021. Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health. *Frontiers in Artificial Intelligence*, 3.

Gervasi, S. S.; Chen, I. Y.; Smith-McLallen, A.; Sontag, D.; Obermeyer, Z.; Vennera, M.; and Chawla, R. 2022. The Potential For Bias In Machine Learning And Opportunities For Health Insurers To Address It: Article examines the potential for bias in machine learning and opportunities for health insurers to address it. *Health Affairs*, 41(2): 212–218.

Gölz, P.; Kahng, A.; and Procaccia, A. D. 2019. Paradoxes in fair machine learning. *Advances in Neural Information Processing Systems*, 32.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Huang, C.; Huang, L.; Wang, Y.; Li, X.; Ren, L.; Gu, X.; Kang, L.; Guo, L.; Liu, M.; Zhou, X.; Luo, J.; Huang, Z.; Tu, S.; Zhao, Y.; Chen, L.; Xu, D.; Li, Y.; Li, C.; Peng, L.; Li, Y.; Xie, W.; Cui, D.; Shang, L.; Fan, G.; Xu, J.; Wang, G.; Wang, Y.; Zhong, J.; Wang, C.; Wang, J.; Zhang, D.; and Cao, B. 2021. 6-month consequences of COVID-19 in patients discharged from hospital: a cohort study. *The Lancet*, 397(10270): 220–232.

Kamiran, F.; and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1): 1–33.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

Lu, J.; Sattler, A.; Wang, S.; Khaki, A. R.; Callahan, A.; Fleming, S.; Fong, R.; Ehlert, B.; Li, R. C.; Shieh, L.; et al. 2022. Considerations in the reliability and fairness audits of predictive models for advance care planning. *Frontiers in Digital Health*, 4: 943768.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, 3384–3393. PMLR.

Mallat, S. 2016. Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065): 20150203.

McNeely, C. L.; Schintler, L. A.; and Stabile, B. 2020. Social determinants and COVID-19 disparities: Differential pandemic effects and dynamics. *World Medical & Health Policy*, 12(3): 206–217.

Now, A. 2021. AI Now Institute. https://ainowinstitute.org/. Accessed: 2023-07-28.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On Fairness and Calibration. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Rajkomar, A.; Hardt, M.; Howell, M. D.; Corrado, G.; and Chin, M. H. 2018. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12): 866–872.

Raza, S. 2022. A machine learning model for predicting, diagnosing, and mitigating health disparities in hospital readmission. *Healthcare Analytics*, 2: 100100.

Raza, S.; Reji, D. J.; and Ding, C. 2022. Dbias: detecting biases and ensuring fairness in news articles. *International Journal of Data Science and Analytics*, 1–21.

Raza, S.; and Schwartz, B. 2023. Constructing a disease database and using natural language processing to capture and standardize free text clinical information. *Scientific Reports*, 13(1): 8591.

Sikstrom, L.; Maslej, M. M.; Hui, K.; Findlay, Z.; Buchman, D. Z.; and Hill, S. L. 2022. Conceptualising fairness: three pillars for medical algorithms and health equity. *BMJ health & care informatics*, 29(1).

Thomasian, N. M.; Eickhoff, C.; and Adashi, E. Y. 2021. Advancing health equity with artificial intelligence. *Journal of public health policy*, 42: 602–611.

Ting, D. S. W.; Cheung, C. Y.-L.; Lim, G.; Tan, G. S. W.; Quang, N. D.; Gan, A.; Hamzah, H.; Garcia-Franco, R.; San Yeo, I. Y.; Lee, S. Y.; et al. 2017. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *Jama*, 318(22): 2211–2223.

Tramer, F.; Atlidakis, V.; Geambasu, R.; Hsu, D.; Hubaux, J.-P.; Humbert, M.; Juels, A.; and Lin, H. 2017. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, 401–416. IEEE.

Wesson, P.; Hswen, Y.; Valdes, G.; Stojanovski, K.; and Handley, M. A. 2022. Risks and opportunities to ensure equity in the application of big data research in public health. *Annual Review of Public Health*, 43: 59–78.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.