Contextual Design Considerations for Fairer Machine Learning in Digital Health Informatics

SEAMUS RYAN, School of Computer Science and Statistics, Trinity College Dublin, Ireland
GAVIN DOHERTY, School of Computer Science and Statistics, Trinity College Dublin, Ireland

The processes of diagnosis, treatment, and support of people dealing with mental health issues are increasingly becoming augmented by Digital applications. These healthcare applications can generate large volumes of data regarding the user including their state and their behaviour. This data, in turn, can be used to create statistical models with the goal of improving the quality of healthcare, for example via outcome prediction or pattern recognition. However, these models are vulnerable to replicating and perpetuating problems of inequality and bias if created using historic data sets or sets cultivated without careful consideration to demographic representation. Proposed methods of quantifying these biases have included algorithmic definitions of fairness. While consideration of fairness should include the societal context a decision is made in, and the practicalities of implementation in the real world, these data and measurement-based approaches are important tools for the evaluation of models. In this paper we consider a number of contextual questions that may be asked by a designer when using algorithmic fairness definitions within the development of digital health applications. We suggest directions for additional research to support the responsible development and adoption of Machine-learning augmented Healthcare tools.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**;

ACM Reference Format:
Seamus Ryan and Gavin Doherty. 2022. Contextual Design Considerations for Fairer Machine Learning in Digital Health Informatics. In *CHI '22 Workshop: Grand Challenges for Personal Informatics and AI, April 30– May 06, 2022, New Orleans, LA.* ACM, New York, NY, USA, 4 pages.

1 MACHINE LEARNING AND DIGITAL MENTAL HEALTH

Machine Learning (ML) has been considered as a potential tool in numerous healthcare environments [3, 8, 10]. Within mental health, researchers have begun to explore the value ML techniques can have for patients [15]; Topics addressed have included diagnostic guidance in clinical depression [12], identification of the most impactful healthcare supporter messages [4], identification of patients off track for treatment progression [5], momentary interventions [2], and the prediction of negative emotions among otherwise healthy users [6].

ML models are built using historical and/or generated data. If this data contains examples of discriminatory decisions, either through mistakes or missing cohorts, then the model created will be discriminatory. If the data does not accurately represent a full range of patient demographics, then the baseline of what is considered normal will be skewed towards the attributes of the over-represented group. This, as well as other technical challenges such as overfitting already demographically skewed data [13], can result in models that perform significantly worse for minority groups when in real healthcare scenarios.

Manuscript submitted to ACM

2 CONSIDERATIONS WHEN DESIGNING FOR FAIRNESS IN HEALTHCARE

ML research has attempted to quantify the fairness of equity and equality numerically. Problems of unequal error rates (an equity issue) and unequal outcomes between demographic groups (an equality issue) have been quantified in terms of model performance and results, respectively. Using these concepts, the ML literature contains precise definitions for fairness, stemming from statistical parity, causal reasoning, and population demographics comparisons [16]. These definitions approach the problem from different perspectives, attempting to ensure inequality is not created via the implementation of the ML models.

Fairness is frequently thought about as an act of retrospective analysis or an auditing step [14]. However, some Fairness goals require consideration at every step of an engineering project. As the act of data gathering, categorisation, and use in model creation has an impact on the long-term assessment of fairness, definitions are fundamental to the implementation process. Modern application engineering involves the frequent iteration and re-deployment of models to a production environment [9] with complex adoption concerns [11]. To keep pace with this, the use of fairness definitions can be evaluated on an ongoing and iterative basis and be assessed as part of the broader approach a team takes to ensure the end-users design expectation are met.

We break down these considerations into two categories, firstly as we focus on designing for the user and their experience by **analysing the scenario** that may be automated as part of an ML model. Once the specifics of the fairness implications of a scenario is understood and there is clarity on how fairness is defined for users, the next step is to understand the **implications of these definitions**.

2.1 What designers may need to understand about their scenario

As discussed, the considerations for fairness are going to be domain and implication specific. As such the importance of comprehending the complete scenario being implemented including the users' expectations is an important starting point.

> ***When to analyse and who is vulnerable*** - Choosing when and who to include in the analysis of fairness are two large and fundamental questions.
> Regarding when, there is a classic dichotomy; do you prioritise a fair process, focusing on ensuring everyone is treated the same way by the application, or a fair outcome, focusing on ensuring all those involved reach the same goal. This ethical decision becomes a design consideration when using ML. The designer needs the tools and language to understand their users' expectations and for fairness.
> Regarding who, a designer will want to consider what demographic groups should or should not be treated the same. For example, in gender-specific support, it may be appropriate to treat gender cohorts differently, in the allocation of preventive care it may not be.
> ***Are there multiple appropriate definitions*** - There will be times during the design of healthcare applications where there may be multiple users who are affected by an ML model's decision. For example, the fairness of healthcare staff, who are having their time allocated by an ML model can be considered distinctly from the fairness of those patients receiving time.
> ***What is the correct measurement***- Measurement Modelling is a quantitative social sciences approach to choosing an appropriate way to measure an outcome [7]. For example, optimising for fairness in 5-year healthcare outcomes would lead to a different model than optimising for more immediate goals like treatment adherence.

What do the statistical approaches mean for users - Definitions grounded in the ML evaluation fields commonly use statistical evaluation terms derived from four key measures within the model. These are false positive, false negative, true positive, and true negative. For all models the theoretical goal is to maximise correct decisions and minimise mistakes. The method used for optimisation usually involves the intentional trade-off of one or more of these metrics. The choice to emphasise one over the others comes with a complex set of assumptions that may not be immediately obvious.

2.2 What designer could consider when reviewing the definition(s)

Once consideration has been given to the scenario, a designer will have a list of, potentially very broad, requirements based on how fairness concerns may impact on users. This list of requirements may also include existing possible fairness issues. The next step may be to refine these fairness definitions and issues to create a more practical set of implementable Fairness metrics.

Are the definitions a requirement or an objective - In a scenario where multiple definitions are implemented for the same model, there will be an inevitable hierarchy of those that are prioritised over others. It can be helpful to consider which definitions are requirements, be they ethical or legal. Others can be considered objectives, where it is desirable to meet that definition but it is not essential to the models' success. For example, in some situations, we may view equal healthcare outcomes as a requirement and equal allocation of healthcare resources as an objective; In a perfect world, both would be met simultaneously but in practice one is a higher priority.

Are any of the chosen definitions explicitly contradictory - Some definitions of fairness are contradictory. The clearest example is a definition that focuses on the exclusion of data (Fairness through unawareness) being contradictory to definitions that required the analysis of results based on that data (Statistical Parity). Other definitions while not directly contradictory can be difficult to satisfy simultaneously. This tension may need to be considered during design.

Can you meet the requirements & assumptions of definitions - Definitions for fairness have inherent requirements and assumptions. One of the most commonly overlooked requirements is the availability of demographic information to be able to assess whether statistical definitions have been met within the models' decision-making process [1]. While in some areas this is possible, data minimisation strategies implemented in the context of sensitive and high-stakes decision-making, including healthcare, mean that demographic variables are often explicitly omitted.

3 FUTURE DIRECTIONS IN FAIRNESS

As discussed in the preceding sections, assessing fairness is a complex and longitudinal process with numerous potential outcomes and no singular correct answer. Designers need additional tools, language, and techniques to be able to meet the expectations of a fair ML system. They will want to be able to sensitise themselves to the fairness concerns of their end-users and stakeholders, extrapolate those concerns into a coherent set of system requirements, investigate when those fairness issues arise throughout the life-cycle of an ML application, and mitigate them once they are detected. Strategies for bias mitigation, be it data augmentation, system redesign, or wider sociological changes, are an important part of building ethical socio-technical systems; A clearer idea of what sort of fairness is most appropriate and how best to implement it is an important early step towards this goal.

ACKNOWLEDGMENTS

This work has been supported in part by Science Foundation Ireland under Grant number 18/CRT/6222.

REFERENCES

[1] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What We Can't Measure, We Can't Understand. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 249–260. https://doi.org/10.1145/3442188.3445888 arXiv:2011.02282

[2] Andreas Balaskas, Stephen M. Schueller, Anna L. Cox, and Gavin Doherty. 2021. Ecological momentary interventions for mental health: A scoping review. *PLOS ONE* 16, 3 (03 2021), 1–23. https://doi.org/10.1371/journal.pone.0248152

[3] Jonathan H Chen, Steven M Asch, and Palo Alto. 2018. Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. *N Engl J Med* 376, 26 (2018), 2507–2509. https://doi.org/10.1056/NEJMp1702071.Machine

[4] Prerna Chikersal, Danielle Belgrave, Gavin Doherty, Angel Enrique, Jorge E Palacios, Derek Richards, and Anja Thieme. 2020. Understanding Client Support Strategies to Improve Clinical Outcomes in an Online Mental Health Intervention. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376341

[5] Jaime Delgadillo, Kim de Jong, Mike Lucock, Wolfgang Lutz, Julian Rubel, Simon Gilbody, Shehzad Ali, Elisa Aguirre, Mark Appleton, Jacqueline Nevin, Harry O'Hayon, Ushma Patel, Andrew Sainty, Peter Spencer, and Dean McMillan. 2018. Feedback-informed treatment versus usual psychological treatment for depression and anxiety: a multisite, open-label, cluster randomised controlled trial. *The Lancet Psychiatry* 5, 7 (7 2018), 564–572. https://doi.org/10.1016/S2215-0366(18)30162-7

[6] Galen Chin-Lun Hung, Pei-Ching Yang, Chia-Chi Chang, Jung-Hsien Chiang, and Ying-Yeh Chen. 2016. Predicting Negative Emotions Based on Mobile Phone Usage Patterns: An Exploratory Study. *JMIR Research Protocols* 5, 3 (2016), e160. https://doi.org/10.2196/resprot.5551

[7] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 375–385. https://doi.org/10.1145/3442188.3445901 arXiv:1912.05511

[8] Steven Lemm, Benjamin Blankertz, Thorsten Dickhaus, and Klaus Robert Müller. 2011. Introduction to machine learning for brain imaging. *NeuroImage* 56, 2 (2011), 387–399. https://doi.org/10.1016/j.neuroimage.2010.11.004

[9] Silverio Martinez-Fernandez, Anna Maria Vollmer, Andreas Jedlitschka, Xavier Franch, Lidia Lopez, Prabhat Ram, Pilar Rodriguez, Sanja Aaramaa, Alessandra Bagnato, Michal Choras, and Jari Partanen. 2019. Continuously Assessing and Improving Software Quality with Software Analytics Tools: A Case Study. *IEEE Access* 7 (2019), 68219–68239. https://doi.org/10.1109/ACCESS.2019.2917403

[10] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C. Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona J. Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher J. Kelly, Dominic King, Joseph R. Ledsam, David Melnick, Hormuz Mostofi, Lily Peng, Joshua Jay Reicher, Bernardino Romera-Paredes, Richard Sidebottom, Mustafa Suleyman, Daniel Tse, Kenneth C. Young, Jeffrey De Fauw, and Shravya Shetty. 2020. International evaluation of an AI system for breast cancer screening. *Nature* 577, 7788 (2020), 89–94. https://doi.org/10.1038/s41586-019-1799-6

[11] Camille Nadal, Corina Sas, and Gavin Doherty. 2020. Technology Acceptance in Mobile Health: Scoping Review of Definitions, Models, and Measurement. *Journal of Medical Internet Research* 22, 7 (jul 2020), e17256. https://doi.org/10.2196/17256

[12] Anastasia Pampouchidou, Panagiotis G. Simos, Kostas Marias, Fabrice Meriaudeau, Fan Yang, Matthew Pediaditis, and Manolis Tsiknakis. 2017. Automatic Assessment of Depression Based on Visual Cues: A Systematic Review. *IEEE Transactions on Affective Computing* 10, 4 (2017), 445–470. https://doi.org/10.1109/TAFFC.2017.2724035

[13] Yubin Park and Joyce C. Ho. 2021. Tackling Overfitting in Boosting for Noisy Healthcare Data. *IEEE Transactions on Knowledge and Data Engineering* 33, 7 (jul 2021), 2995–3006. https://doi.org/10.1109/TKDE.2019.2959988

[14] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 33–44. https://doi.org/10.1145/3351095.3372873 arXiv:2001.00973

[15] Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine Learning in Mental Health. *ACM Transactions on Computer-Human Interaction* 27, 5 (10 2020), 1–53. https://doi.org/10.1145/3398069

[16] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*. ACM, New York, NY, USA, 1–7. https://doi.org/10.1145/3194770.3194776