# Medical Machine Learning technologies as an example for necessary ethical trade-offs in ML

Albrecht, Thomas (5733587) Petruck, Julian (3857386)
Jaques, Arthur (5998179) Voulgaris, Sotiris (6013093)

February 2, 2022

**Abstract**

We use the application of Machine Learning to healthcare as a case study of ethical trade-offs. We concentrate on trade-offs between privacy and predictability in the use of patients' data, between group fairness and individual fairness in the attempt to make ML-based systems "fair", and between fairness and prediction accuracy when applying fairness constraints to the ML systems. Firstly, we examine and discuss whether those trade-offs are unavoidable, and relate them to moral dilemmas in moral philosophy. Secondly, we examine the results that are obtainable with regards to those trade-offs (where do we want to lie on the Pareto frontier?). In the case of the trade-off between group fairness and individual fairness, we dive into the conflict between the aggregate and the individual, between the population level view of the "average man" and the concrete individuals that are affected by the ethical policies. In our critical analysis, we relate the existing best practices in medicine and their existing literature (as an example, the four principles proposed by Beauchamp and Childress), and the fairness tools and analyses provided by the ML community. As a consequence, we suggest what the communities could learn from each other and what differences need to be resolved.

## Contents

# 1 Introduction

### The intersection of machine learning and medicine

- ML is great, blah blah, historical setting, current application.
- Fairness problems identified in the literature.
- Trade-offs and moral dilemmas: algorithmic, philosophical, medical view.
- Peculiarities of health care as application field: inevitable moral dilemmas, impossibility of the "do nothing" solution, developed moral literature, high stakes, less readiness to sacrify performance, human comparisons. true predictors: "difference does not always entail inequality. In some instances, it is appropriate to incorporate differences between identities because there is a reasonable presumption of causation [1, e221]"

### Research questions

- (Where) Are trade-offs necessary? Are algorithmic trade-offs and moral dilemmas different?
- What are the current results in ML? Are they going in the right direction?
- How are trade-off situations currently handled in medical practice? What are hidden questions?

- Are fairness problems of ML applied to medicine new problems intrinsic to the technology, or are they inherent to medical practice?
- Is "doing nothing" really an acceptable solution?
- Can we implement biomedical principles in ML?
- What can ML learn from medical ethics?

**Methods and sources**

- Literature search from different sources: philosophy, medical ethics, fair-ML.
- Theoretical reflections and linking literature sources and fields; interdisciplinary connections.
- Work does not propose concrete implementation solutions.

# 2 Trade-offs in machine learning and automated systems

- We identify 3 sources of ethical discussion...
- Why not concentrate on one? All must be approached when the solution is implemented.
- Must be considered together, since they are not orthogonal axes. Eg., privacy might mean reducing the individual even more to group characteristics. Additionally, privacy and fairness might be in conflict: targeted data collection to correct data biases "may pose ethical and privacy concerns as a result of additional surveillance" [2, p. 8].

## 2.1 Privacy and predictability

# TODO: specify content Non-maleficience, trade-off [3].

### 2.2 Group fairness and individual fairness

- Why we think Binns 2020 does not cancel the problem. cp. "Given the epistemic uncertainty surrounding the association between protected identities and health outcomes, the use of fairness solutions can create empirical challenges" [1, e221]. negative legacy, labeling prejudice, sample selection bias [2, p. 6].
- Specificity of medicine: groups sometimes DO matter in the prediction. "difference does not always entail inequality. In some instances, it is appropriate to incorporate differences between identities because there is a reasonable presumption of causation" [1, e221] Importance of the "causal structure between latent biological factors such as ancestry and their associated diseases across ethnic subpopulations" [2, p. 3].
- ML systems have the (demonstrated in practice) potential to discriminate, even if group information is not included, through for example leakage of ethnicity, which is then used as a shortcut to make the predictions (reproducing, or even amplifying, historical bias) [2, p. 3]. For this reason, so-called fairness through unawareness is insufficient in non-discrimination. [2, p. 5].

### 2.3 Fairness and prediction accuracy

- Specificity of medicine: allocation of physical benefits and harms. Non-maleficience?
- "difference between an idealised model and non-ideal, real-world behavior affects metrics of model performance (eg, specificity, sensitivity) and clinical utility in practice." [3, e221].

## 3 Trade-offs in medicine

#Goal: identify ethical issues and trade-offs pre-existing the application of ML, describe what principles are used in deciding for the best solution, examine how they are dealt with currently. Healthcare disparities are a well-accepted reality, and "often encompass all 5 domains of the social determinants of health as defined by the US Department of Health and Human Services (economic stability, education access and quality, healthcare access and quality, neighborhood and built environment, and social and community context)" [2, p. 2].

### 3.1 Trade-offs in unassisted medicine

Limited resources, uncertainty, ... "Trade-off" is a tendentially mathematical notion. Hence, mentioning trade-offs in the field of medicine or ethics might cause defensive reactions, because of the supposed complexity of ethical problems. Suggesting that doctors apply trade-offs in their practice is a contestable affirmation, since the nature of their ethical deliberations is necessarily partly non-mathematical. Hence, a more appropriate term here is ethical (or moral) dilemma, which is a problem that arises when opposing values or principles co-occur [3, p. 351]. Fundamentally, however, trade-offs and practical solutions to moral dilemmas are the same thing: a decision on how much to respect principles that can not be fully respected at the same time.

### 3.2 Principles of medical bioethics

A good starting point for ethical discussions in medicine are the well-established guiding principles in biomedical ethics proposed by Beauchamp and Childress: respect for autonomy, beneficence, non-maleficence and justice [3, pp. 344-345], [4, p. 2], [5, p. 2].

Describe, related to ML. Focus on beneficence vs non-maleficence.

### 3.3 Pragmatism

A solution has to be found, since non-action is worse than everything. In the face of uncertainty, leeway is left

## 4 Combining machine learning and medicine

### 4.1 Old and new problems

As discussed in the previous sections, the practical problem of applying Machine Learning to health care tasks carries with it a certain number of unavoidable decisions about the relative importance of contrasting principles. In short, trade-offs have to be made. In this essay, we concentrate on trade-offs related with fairness considerations. An interesting aspect of such trade-offs is their origin: what are their causes? We argue that most, if not all, fairness-related trade-offs originate from the decision (medical) problem itself, and not from the technology used to solve it. Look for resources in Bob's technology lecture. That is, Machine Learning does not introduce additional fairness problems, but simply specifies existing trade-offs and makes them unavoidable.

As a concrete example, take the problem of unbalanced training data causing the Machine Learning algorithm to reach better performance on over-represented groups. This problem is hidden, but still present, in unassisted medicine. Medical practitioners arguably learn the best treatments partly from experience. If the majority of their patients come from a particular group, it is very likely that they will "overfit" their knowledge to that group, or at least be able to predict their response to particular treatments better. Similarly, textbook knowledge is derived from observations from medical practitioners and/or statistical studies. Biased data informing those studies will bias the observed results. This effect is exemplified by hearth health research, where research on biased data (higher proportion of men) lead to uneven improvements in heart attacks treatment with respect to gender [1, e221]. The data that informs that knowledge is probably very similar to the data used to train ML algorithms. The resulting inferences will hence be similarly biased as a consequence of data imbalances. In this sense, ML systems even have a better potential to solve the problem, using for example importance weighting or under-/over-sampling [2, pp. 6-8]. By no means do we intend to suggest that the solution is easy, since blindly applied technical fixes may introduce undetected harms (contrasting with the bioethical principle of non-maleficence). However, ML has the potential to fix biases in a way that traditional medical practice can not [1, p. e222] (although traditional statistics can, see e. g. importance sampling).

Similarly, the problem of data privacy pre-exists the entry of Machine Learning in the medical field [3, p. 346]. Privilege bias (models being developed for diseases that disproportionately affect a certain group) [5, p. 5] is a problem that exists in classical statistical studies as well cite Jackson's article in the Guardian. In short, the trade-offs we analyzed (group fairness and individual fairness, privacy and predictability, fairness and predictability) exist independently of the system used to make decisions. That is, they are not inherent to the technology used to solve them, ML, but to the goals and requirements of the system. ML can be used to proactively advance health equity (beneficence), and not only avoiding harms (non-maleficence) [5, p. 2].

We do not, however, argue that ML does not introduce any new ethical problems, but that they are not fairness-related. For example, ML systems applied at a large scale unify decisions, exponentially increasing the impact of failures. Less federated decisions make for less error-robust systems, and unified treatment and strategy gives fewer indications about what works well, potentially reducing the possibilities for learning from single experiences. Additionally, automated decision systems distribute the responsibility for the decisions the system takes, making it very difficult to attribute responsibility for potential misjudgments [4, p. 6]. If the decisions are not taken automatically, the practitioner might still rely too much on them and avoid a right call that would contradict the suggestion of the ML system [4, p. 4] (automation bias, [5, p. 4]). Additionally, they might pay less attention to the decisions that are assisted by technology. The lack of explainability of the decisions taken by automated systems might additionally contrast with the biomedical principle of the respect of autonomy, since it reduces the patient's possibility to exert informed consent [3, p. 346]. Finally, the trust of the public in the decisions taken by automated decision systems may be low, making their large-scale use politically difficult [4, p. 4].

### 4.2 Clarification of problems

Main point: the mathematical rigor of ML forces us to think about those problems; this is a positive feature and not a disadvantage. How can ML actively help advance health equity and fairness? Firstly, it imposes the need for precise definitions of what is meant by terms like "discrimination", "equity", and so on. Secondly, it forces the developers of the system to choose precise weights for the principles that they want to respect, and explicitly accept the existence of trade-offs that are inherent to the problem. Thirdly, it makes the goals and evaluation metrics (and their implied definition of what a "good" solution looks like) clear. Knowing that those goals influence the results, with for example pure efficiency potentially leading to the propagation of health inequities [5, p. 2], the importance of each objective has to be decided upon (and hence, the chosen position on the Pareto frontier).

### 4.3 Advantage of inaction

- Positive versus negative harms: in doubt, do nothing.

- This reasoning is much harder to apply to critical problems as those emerging in medicine.

# 5 Trade-off part

## 5.1 Tradoffs Intro/in general

inherently contradictory(trival for conflict as opposed to view that tradeoffs dont exist or should be avoided)

### 5.1.1 Pareto Efficiency

- Multi-objective optimization

## 5.2 Tradeoffs in Medicine/Ethics

### 5.2.1 Patient-Tradeoff

## 5.3 Tradoffs in ML/fairness

### 5.3.1

## 5.4 Beyond Pareto

# 6 Conclusions

Answer (again), one after the other in separated paragraphs, the research questions stated in the introduction. Examples: Are trade-offs an inherently technical problem? When is (in)action justified? ...

## References


[1] Melissa D McCradden, Shalmali Joshi, Mjaye Mazwi, and James A Anderson. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health*, 2(5):e221–e223, 2020.

[2] Richard J Chen, Tiffany Y Chen, Jana Lipkova, Judy J Wang, Drew FK Williamson, Ming Y Lu, Sharifa Sahai, and Faisal Mahmood. Algorithm fairness in ai for medicine and healthcare. *arXiv preprint arXiv:2110.00603*, 2021.

[3] Marianne WMC Six Dijkstra, Egbert Siebrand, Steven Dorrestijn, Etto L Salomons, Michiel F Reneman, Frits GJ Oosterveld, Remko Soer, Douglas P Gross, and Hendrik J Bieleman. Ethical considerations of using machine learning for decision support in occupational health: An example involving periodic workers‘ health assessments. *Journal of Occupational Rehabilitation*, 30:343–353, 2020.

[4] Jessica Morley, Caio CV Machado, Christopher Burr, Josh Cowls, Indra Joshi, Mariarosaria Taddeo, and Luciano Floridi. The ethics of ai in health care: A mapping review. *Social Science & Medicine*, 260, 2020.

[5] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.