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 26, 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. Allocation of positive goods might be different from prevention of harm, eg in healthcare settings. Idea: harm distribution is different from benefit distribution. Trying to ensure equal harms in a setting where medicine can very well solve one group's problems seems illogical. 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 [4, 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 part

2.1 Trade-offs Intro/in general

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

The notion of a tradeoff describes a decision between multiple (usually mutually contradictory) objectives, in the sense that a gain in one objective results in loss in one or more other objectives. The concept of trade-offs is very common in everyday life, ...

...

But in economics especially, trade-offs play a huge role and there have been multiple approaches to formalize them.

2.1.2 First derive mathematical formulation used in economics(pareto efficiency and multi objective optimization)

To approach choosing a feasible decision (allocation) for various types of trade-off we will introduce multi-objective optimization. A general multi-objective optimization problem can be written in the following way:

$$min(u_1), \quad s.t. x \in X$$

Here X denotes the set of all feasible decisions and $u_i(x)$ the utility/objective function for each dimension. For a non-trivial multi-objective optimization problem it is not possible to minimize every single objective function at the same time. Thus the notion of Pareto optimality is introduced: A decision $x \in X$ is said to Pareto dominate another solution $x' \in X$ if the following both hold:

1. $u_i(x) \leq u_i(x'), \forall i \in 1, 2, \ldots, k$
2. $u_j(x) < u_j(x'), \forall j \in 1, 2, \ldots, k$

Such a decision is also called Pareto optimal/Pareto efficient. The set of all Pareto optimal decisions is called the Pareto front. If the optimization problem is two-dimensional the Pareto front can be visualized in an intuitive way: The objectives are the axes in a 2D plane, moving along the Pareto front showcases how increasing one objective decreases the other one[maybe add a figure here with example caption (accuracy fairness seems good)]. Note that Pareto optimality does not ensure a morally right decision in any sense (change this).. . .

But the concept of Pareto optimality alone usually won't actually leave us with a single optimal or "best" answer to our decision problem. Rather, the approach eliminates all "strictly worse" possible decisions in the feasible set and the decision maker is faced with a new problem. To now choose one solution from the Pareto front multiple methods can be used. Depending on the problem at hand the decision maker could (or rather has to) potentially incorporate additional prior information (knowledge/preference).

. . .

2.1.3 Talk about tradeoffs in medicine/ethics in general

As already hinted to earlier, tradeoffs play a central role in medicine. They can appear in different areas in the field, be it at high level healthcare policy decisions or when considering treatment options for one specific patient. Sometimes those tradeoffs arise when the decision involves a moral or ethical dilemma: If for example the administration of a treatment could harm

the patient as a side effect one might still choose to treat the disease if it is the lesser evil (e.g. chemotherapy[cite]). In this case the the physician faces multiple tradeoffs: He has to consider the effectiveness of the administered treatment (which is uncertain for the given patient) the likelihood and magnitude of possible adverse side-effect (which are also uncertain for the given patient). He also has to consider ...

By applying the concept of Pareto optimality one could even say that any decision that doesn't involve a tradeoff of some sort would be trivial to make, because it would have a unique maximum. Of course it would be desirable to avoid many of the tradeoffs in the sense of maximizing all objectives simultaneously, but that maximizing solution might not be in the feasible set, i.e. a possible decision at the given time.

...

2.1.4 elaborate on tradeoffs in medicine and whether the mathematical formulation can/should be employed (it shouldn't be)

2.2 Beyond Pareto

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)" [1, p. 2].

3.1 Trade-offs in unassisted medicine

Limited resources, uncertainty, ...

"Trade-off" is an originally 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 [7, 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 [7, pp. 344-345], [5, p. 2], [6, 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 many 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 often 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 [4, 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 [1, 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 [4, 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 [7, p. 346]. Privilege bias (models being developed for diseases that disproportionately affect a certain group) [6, p. 5] is a problem that exists in classical statistical studies as well [3]. 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) [6, 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 [5, 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 [5, p. 4] (automation bias, [6, p. 4]). Additionally, they might pay less attention to the decisions that are assisted by technology. The complex models of ML and the large amount of features used help make more personalized decisions for the individual, but decrease their understandability [6, p. 4]. 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 [7, 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 [5, 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 [6, p. 2], the importance of each objective has to be decided upon (and hence, the chosen position on the Pareto frontier). To summarize these advantages, we can say that ML, despite the typical complaints about its inscrutability, in a way helps enforce the transparency of the decisions taken. This is an important factor especially when comparing their use to current practice and human-centered decisions, where the practitioner's values are necessarily at least indirectly influencing their decisions, probably without being stated precisely.

4.3 New trade-offs

The use of ML as assisting systems rather than replacements of clinicians altogether complicates the discussion about biases further. The end effect of the integration of ML tools in medical practice is a complex function of the interaction of their results and their usage by clinicians on patients [6, p. 4]. Julian's paragraph here

4.4 Advantage of inaction, and the sin of perfection

- Positive versus negative harms: in doubt, do nothing.
- This reasoning is much harder to apply to critical problems as those emerging in medicine.

The visceral resistance to the use of any technological system that shows any behavior deemed as unjust might be stopping improvements in overall care, and can be considered problematic. Do we want, for example, to refute to apply any system that does not lead to equalized outcomes [6, p. 5] but only equalized benefit [6, p. 5]? How do we justify keeping the unfair status quo by avoiding solutions that would improve care in general and stratified across sensitive groups, just because those solutions do not perfectly solve the problem?

Since our view of the world is partial and hence stochastic, we have to accept that any decision is subject to uncertainty and to the possibility of being 'wrong'. The empiricist's answer to this problem is to observe the effects of the decision and adapt his assumptions and knowledge based on them. The advantage of actively trying a solution, despite the uncertainty about its results, is the positive feedback loop that it creates. If we observe the development of ML systems under this lens, we can accept that solutions will evolve over time based on the results they get (see lecture 12, slide 5). That is, the fact that we change the way decisions are made will change the underlying data distribution and offer us more insights about the real sources of group differences. For example, actively trying to correct for historical bias by applying equal allocation principles [6, p. 6] will give us more diverse data based on which to infer the causes of past differences, and what the best approach is to solve them. A possible solution is hence to develop system that we deem the more appropriate with the current knowledge, accept the imperfection and improve them over time as they get results. ML systems are not tools that once applied will remain forever the same: they should be closely monitored and improved over time [6, p. 7]. However, the dynamics of the entire ecosystem make it very difficult to predict its evolution. Furthermore, very little work has been done in ML to assess the evolution of the data distribution when decisions are taken by ML systems adjusted for fairness. Economics literature in affirmative action may be helpful in analyzing the problem [2].

Data 'set' notion, necessity to accept bias in data, data as a time-dependent snapshot, evolution, effects of interventions, dynamics, ...

Examine the idea that "clinicians bury their mistakes" vs aviation, ... for why it may seem more problematic in the field. Culture of improvement.

5 Conclusions

Are trade-offs an inherently technical problem? When is (in)action justified? ...

6 Arthur's ideas

6.1 tradeoffs, or maybe introduction?

multiple objectives motivation We identify three axes of conflict when implementing fairness into ML systems. Firstly, assuring privacy requires modifying the data (thus removing information), which probably leads to a deterioration of prediction accuracy. (cite e.g.https://www.nature.com/articles/d41586-020-02454-7). Non-maleficience, trade-off [7].Secondly, the typical implementation of fairness into ML systems is done in the form of group fairness measures, i.e., requires the separation of people into groups, usually by so-called sensitive attributes. This leads to a conflict between individual fairness, with individuals wishing to be judged independently of their group identity, and group fairness, which tries to correct for supposed historical and data biases. It further raises constraints of group belonging and typicality (is it advantageous to be "average" in its own group?). (Discuss Binns). Use slideset 9, slide 22 for Binns comment. Individual justice ideas seem to go exactly in the opposite direction of "Explainable AI", since they basically say that concepts that can not be put into words should be used to base a decision. In general, Explainable AI requirements contrast with "AI cannot make human-like judgements". The elements to take into account when deciding on what metric of fairness to use are multiple. On the one hand, we need to decide what moral principles we want to follow, i. e. what we intend by equal or just treatment. What do we consider distributive justice? What is the resource that has to be distributed? Do we care about the end-result, or only about promising equal expectancies? On the other hand, we have to provide a model about the sources of unfairness in the data and model we use. In ML terms, we have to state our assumptions about the data-generating process. For example, assuming historical bias means putting into question the validity of the training labels, and hence accuracy on them as a performance measure [6, p. 6].

- 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" [4, e221]. negative legacy, labeling prejudice, sample selection bias [1, 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" [4, e221] Importance of the "causal structure between latent biological factors such as ancestry and their associated diseases across ethnic subpopulations" [1, 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) [1, p. 3]. For this reason, so-called fairness through unawareness is insufficient in non-discrimination. [1, p. 5].

Thirdly, transforming the objective from a single objective of performance to a multiple objective of performance and fairness leads to in general worst performance. We thus arrive at a trade-off between prediction accuracy (or whatever performance measure is used: sensitivity, specificity) and fairness.

- 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 (e.g., specificity, sensitivity) and clinical utility in practice." [7, e221].

Why not concentrate on one tradeoff? 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" [1, p. 8].

Trade-offs as inevitable features of decision problems on a broader view, trade-offs are the basic problem of human governance. How many resources we allocate for one problem, leaving less for another one. (Almost) every decision has positive and negative effects. So subjectivity is always present, and we all accept (if only implicitly) the existence of trade-offs in every decision.

References

[1] Richard J Chen et al. "Algorithm fairness in AI for medicine and healthcare". In: *arXiv preprint arXiv:2110.00603* (2021).

[2] Alexandra Chouldechova and Aaron Roth. "A snapshot of the frontiers of fairness in machine learning". In: *Communications of the ACM* 63.5 (2020), pp. 82–89.

[3] Gabrielle Jackson. *The female problem: how male bias in medical trials ruined women's health*. Nov. 13, 2019. URL: %5Curl%7Bhttps://www.theguardian.com/lifeandstyle/2019/nov/13/the-female-problem-male-bias-in-medical-trials%7D (visited on 02/02/2022).

[4] Melissa D McCradden et al. "Ethical limitations of algorithmic fairness solutions in health care machine learning". In: *The Lancet Digital Health* 2.5 (2020), e221–e223.

[5] Jessica Morley et al. "The ethics of AI in health care: A mapping review". In: *Social Science & Medicine* 260 (2020). ISSN: 0277-9536. DOI: https://doi.org/10.1016/j.socscimed.2020.113172. URL: https://www.sciencedirect.com/science/article/pii/S0277953620303919.

[6] Alvin Rajkomar et al. "Ensuring fairness in machine learning to advance health equity". In: *Annals of internal medicine* 169.12 (2018), pp. 866–872.

[7] Marianne WMC Six Dijkstra et al. "Ethical considerations of using machine learning for decision support in occupational health: An example involving periodic workers' health assessments". In: *Journal of Occupational Rehabilitation* 30 (2020), pp. 343–353.