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

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

March 1, 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

Machine Learning (ML) in health care and medicine has grown to be one of the most discussed, but also most promising applications of the ever-growing technology of ML. In recent years, more and more research has shown ML to be an effective way of supporting health care practitioners in a great diversity of ways [17, 21]. However, there is also growing concern about the implications the deployment of ML has for the future of health care and medicine.

While there are ongoing public discussions about ML replacing humans as workers in many different ways [citation needed], many scholars have made clear that ML and AI tools will not replace clinicians in the near future but rather be integrated as support systems, for example as clinical decision support systems (CDSS) [15].

CDSS have been used since the 1980s with growing success [20]. Only in recent years, the involvement of Machine Learning in those systems has led to a new regulatory situation. Still, their deployment and success can tell us a lot about the way to go with ML tools. For example, although closed loop systems, i.e., systems were every step of the process from diagnosis to drug intake is computerized and monitored, do already exist they are not commonly used, partly due to costs but certainly also due to the involved surveillance environment for patients [20].

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 [13, e221]"

We consider CDSS a particularly interesting use case of ML when it comes to fairness and more generally ethical discussion, since the biomedical field historically played an important role when it comes to ethical principles and deliberations. It helped displace purely theoretical ethical deliberations (meta-ethics) to more concrete, unavoidable, and tangible questions (applied ethics). It further helped replace purely relativist, subjectivist, and psychological investigations of ethics [22]. Broad, universal moral principles were replaced by case studies, for example arising in clinical medicine. Finally, ethical considerations started taking into account the roles and relationships of the actors present (for example, by recognizing the authority relationship between doctors and patients). Poetically put, "Medicine saved the life of ethics" [22]. Add other sources

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?

In our investigations, we draw on literature from the fields of fair-ML, economics, medicine, biomedical ethics, and philosophy. We relate fairness problems raised by the Fair-ML community to ethical questions raised in the medical field. We look at well-known ethical principles from the biomedical literature to determine their usefulness in enlightening the ethical discussion about CDSS. We consider the identified moral dilemmas as an issue of trade-offs, and examine economists' work on the topic for a meaningful formalization.

2 Trade-offs part

The notion of a trade-off 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. On a broad view, trade-offs are the basic problem of human governance. How many resources we allocate for one problem, leaving less for another one. Trade-offs are intuitively understood from a young age as they are very common in everyday life, and encompass all human decision-making. Biology, evolutionary theory, and more precisely the human body can be understood in terms of trade-offs [12].

But in economics in particular trade-offs are of special interest, they are a central point of study in the field. Accordingly, economists have proposed multiple approaches to formalize them. One such approach, which is so widespread and commonly used that it can be regarded as a convention, is Pareto efficiency and the Pareto front.

2.1 Multi-objective optimization and Pareto efficiency

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

$$min\ F(x) = (u_1(x), \ldots, u_k(x)), \quad s.t.\ x \in X$$

Here X denotes the set of all feasible decisions and $u_i(x)$ the utility/objective function representing the k dimensions. 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. $\forall\, i \in 1, 2, \ldots, k : u_i(x) \leq u_i(x')$

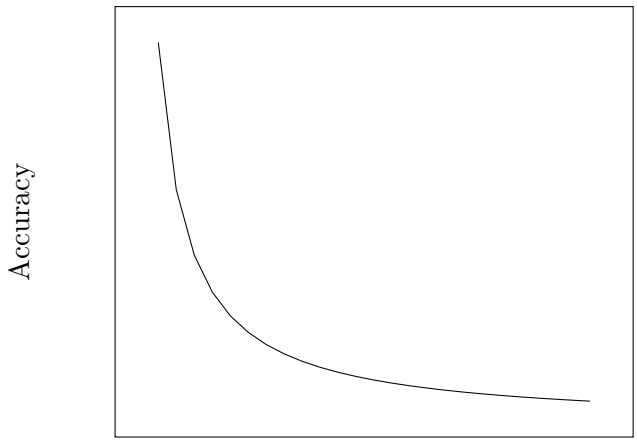

Figure 1: placeholder

2. $\exists\, j \in 1, 2, \ldots, k : u_j(x) < u_j(x')$

Such a decision is also called Pareto optimal or Pareto efficient. Any Pareto optimal decision cannot be further improved for one objective unilaterally without resulting in loss in one or more other objectives. 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

[add a figure here with caption (accuracy fairness)].

Note that Pareto optimality doesn't ensure anything beyond the property derived above. In particular it doesn't provide any guarantees about a "fair" or normative allocation or decision.

...

But the concept of Pareto optimality alone won't result in 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. She now has to choose one solution (decision) from the Pareto front. Depending on the problem at hand the decision maker could (or rather has to) potentially incorporate additional prior information (knowledge/preference).

How to choose along pareto front (how to solve)...

- a priori, incorporating priors like knowledge/preference/ best practices/experience
- a posteriori
- other MCDM approaches

Tuning via "knobs" (lecture 3) is also just travelling along pareto front

2.2 Tradeoffs in fair Machine Learning

As with any technology [cite] there are many tradeoffs inherent in machine learning systems and algorithms.

... (maybe three axes of conflict arthur part)

Here we are going to characterize three fairness related tradeoffs in machine learning systems:

Accuracy vs. Fairness (Cost of Fairness in binary classification) Prediction accuracy is a very desirable property in machine learning systems, maximizing it is often the primary goal of the employed algorithm. Fair machine learning is concerned with identifying and mitigating bias and discrimination of sensitive attributes in ML systems. Ideally we would like to achieve optimal accuracy while not discriminating with respect to any sensitive feature. But as demonstrated empirically in e.g. [9] and [24] avoiding discrimination (or achieving a certain level of fairness) often directly results in the loss of prediction accuracy.

Furthermore, [14] showcases that this tradeoff is a property inherent in the data and doesn't depend on the algorithm used when learning on a modified problem subject to a fairness constraint. Here the accuracy/fairness tradeoff depends on the "alignment" of the label and the sensitive feature, in the sense that if the label and sensitive feature are highly correlated, ensuring a certain level of fairness will result in huge loss of accuracy. Conversely, if the sensitive feature and the label are independent of each other, we can achieve perfect fairness while retaining the full accuracy.

To guarantee a certain amount of fairness, subject to a certain fairness measure, while retaining the maximum accuracy possible under that fairness constraint we can consider the Pareto front of this tradeoff. It is characterized by ... (compare [] and [])

Group vs Individual Fairness

Accuracy vs. Privacy Privacy, just like fairness, is another information based harm [cite lecture?] that, unlike fairness, ...

...

2.3 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.

...

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

3 Trade-offs in medicine

3.1 Trade-offs in traditional medicine

- Limited resources
- Cost vs health care quality trade-off
- Can the individual decide it? (Private vs public insurance).

Mentioning trade-offs in the field of medicine or ethics might cause defensive reactions because of the mathematical flavor they carry, which clashes with 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 citation required. Hence, a more accepted term here is ethical (or moral) dilemma, which is a problem that arises when opposing values or principles co-occur [18, p. 351]. Fundamentally, however, trade-offs and practical solutions to moral dilemmas are the same thing: a decision on how much to weight principles that can not be fully respected at the same time. Far from purely qualitative reasoning, a step in the quantitative dimension of trade-offs is shown for example by evidence-based medicine [12], which serves to inform decisions on what risks are to be taken with the promise of some potential benefit.

The perhaps most obvious trade-off in the practice of medicine, that every doctor understands, is the one between the potential gains and the risked losses [12]. In fact, one can go as far as to "conceptualize medicine itself as the art of managing trade-offs" [12]. From the doctor's allocation of time to specific patients, to the risk of switching to a new potentially better treatment, to the decision of how aggressively to treat terminal patients, every hard decision a medical practitioner has to take entails a trade-off.

While we concentrated above on the decisions the single practitioner has to make, health care itself, as part of governance, is ridden with trade-offs. Health care systems themselves are administrated according to risk-benefit analyses, both as part of the overall governance budget and within the system (which operations to prioritize, what costs to cover, and others) [4].

Furthermore, much attention is paid to preserving privacy when using medical records and clinical data for scientific studies. The European GDPR is for example an important personal data protection law, that because of unclarity and unresolved legal issues often stalls scientific research and progress as a result [5].

3.2 Fairness in traditional medicine

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].

3.3 Solutions from medical ethics

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 [18, pp. 344-345], [15, p. 2], [16, p. 2]. The guiding ideas of biomedical ethics can be used to assess specific applications of ML to health, for example decision support in occupational health, by considering the potential benefits and risks with respect to those principles [18].

Despite the interest of considering biomedical ethics, actual practice seems to indicate that case-by-case evaluations of the moral implications of medical decisions are more useful than principled approaches. Toulmin reports for example how a commission of people from different backgrounds, faced with specific practical problems, were able to reach some consensus (disagreeing at most about the degree of the decisions), all while furiously disagreeing about the principles supporting their decisions [22]. Physicians typically exert their clinical judgment only after collecting a precise case history, instead of following general theoretical considerations early on [22].

4 Combining machine learning and medicine

4.1 Fairness problems at the intersection of ML and medicine

4.2 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 the previous sections, we concentrated on trade-offs related with fairness considerations. An interesting aspect of such trade-offs is their origin. We argue that many fairness-related trade-offs originate from the decision (medical) problem itself, and not from the technology used to solve it. This means that the analysis of fairness is necessarily domain-dependent, and in our case must draw on medical ethics literature. Hence, we suggest abstracting from Machine Learning itself and regard it as a general technology used to solve a pre-existing problem. The ethical considerations will hence be based on both medical ethics (to allow for domain-dependence) and philosophy of technology, recognizing the interaction between our behavior and the technology we use [18]. Our reason to, on a first analysis, disregard the specifics of Machine Learning is that in

many cases, 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 [13, 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 [13, p. e222] (although traditional statistics can help, see e. g. importance sampling).

The issue of reducing an individual to a group identity already exists in statistics, and arises in classical clinical practice as well. Despite claims about the possibility of 'personalized medicine' thanks to ML methods, individuals are still reduced to their features [18]. Although it is true that personalization might be an illusion, ML allows for more precise groupings

Similarly, the problem of data privacy pre-exists the entry of Machine Learning in the medical field [18, p. 346]. Privilege bias (models being developed for diseases that disproportionately affect a certain group) [16, p. 5] is a problem that exists in classical statistical studies as well [8]. 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) [16, p. 2].

4.3 Potential benefits and the sin of perfection

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 [16, 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, by requiring the ethical position to be written down explicitly [23]. 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. Let us reiterate this point: in many cases, ML technologies would simply assist or replace human decision-makers, so the analysis of their impact must be relative to the current human decision-making, and not an absolute decision about whether they act 'perfectly fairly' or not.

- 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 [16, p. 5] but only equalized benefit [16, 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 [16, 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 [16, 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 [3].

Draw on "How humans judge machines."

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.

4.4 New trade-offs

Until now, we argued that most of the problems identified by the Fair-ML community pre-exist the application of Machine Learning to health. However, we can identify some issues that arise specifically from the combination of medicine and Machine Learning: that is, ethical issues that uniquely emerge from this technology [18]. In particular, 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 [16, p. 4]. Julian's paragraph here

We have argued that there are no new fairness trade-offs at the intersection of machine learning and medicine but rather that the preexisting ones

are preserved, increased or possibly decreased. However, the deployment of machine learning methods in the medical context does introduce new trade-offs into medicine apart from the fairness domain. So let us zoom out of this domain to see what is happening when ML and medicine are combined.

ML tools in medicine are often discussed as a human vs. machine situation - where the ML tool outperforms the human they should and in the near future will be substituted. However, making a binary decision out of this does not seem to be the optimal solution. Different studies found that combining AI and human evaluation can achieve better results than either of the two on their own [11, 17, 19, 21]. One of those studies also found that especially for harder cases the assisted accuracy was very high compared to the unassisted accuracy when the ML model's prediction was correct, but that it was also painfully low in cases where the ML model's prediction was incorrect [11]. So instead of a binary decision we are left with a new situation that fits our understanding of trade-offs. How are ML and human evaluation best combined to achieve the optimal accuracy? This might heavily depend on the task at hand. For example, for skin cancer classification where the input is only a cropped image of the potential carcinoma or melanoma, the algorithms decision alone might the enough. However, for identifying diseases in a breast X-ray, a much broader task than skin cancer classification, algorithmic and human judgement might need to be combined for the optimal solution. maybe more AI for less experienced, less AI for more experienced? This is a behavior that was already found for CDSS without ML components, where more experienced doctors were shown to ignore the assistance more often [20].

Often, ML tools only work for specific tools, i.e. detecting one or a couple of diseases in an X-ray. While the accuracy rate here is often high the broadness of the analysis is very limited compared to a doctor [21]. This could be identified as a trade-off between high accuracy with a narrow focus on the one hand and lower accuracy with a broader focus on the other.

The current way of handling medical data differs heavily from the way data is used in ML [6]. Unfortunately, to make ML tools work properly there is a need for huge amounts of data that will be shared with the respective companies and researchers. This creates a trade-off between the classical handling of medical data and a necessary data collection.

A trade-off that is not inherent to the application of ML in medicine and health but that grows to a new importance in this field is between explainability and accuracy/performance [10, 21] (imitations and challenges in [21]) . While explainability plays an important role to foster trust in ML there is probably no other field where this is as important as in medicine and

health care. This can also be related to another, rather philosophical trade-off: If a person is sceptical about using ml on their diagnosis, how can we trade-off a potentially better diagnosis against respecting the persons wish with possibly risking a less accurate or even wrong diagnosis? Explainability might be a decent solution to gain the trust needed, but it might also worsen the accuracy thus actually making the mistrust in the technologie more reasonable.

The developement and deployment of ML tools in medicine and health care will and does already cost a lot of money [6]. At the same time, the health care system in general in countries like the US is heavily underfunded. So much so that live expectancy began to decrease again in the US [21]. Thus, there can be a trade-off identified between the financing of ML tools and the health care system in general. If the huge investments in ML tools will only benefit a small wealthier part of society, those investments are questionable if the health care systems continue to be underfunded. This is even more the case since there are not yet many ML tools ready for clinical application which makes this money an investment into the future while there persist acute issues that would need to be tackled here and now.

Many of the trade-offs discussed can essentially be broken down to one question: How much do we benefit from the use of ML in medicine? What might be bad for us, for example could the digitalization and sharing of our health data lead to misusage by health care providers? If I know that I will most probably benefit from sharing my data on the other hand, I will be more likely to do so. This would create a very general trade-off between benefits and caveats oof ML tools in medicine and health care. ([21] Increased Efficiencies) ([6] Transparency)

Another trade-off exists between the way medical devices are traditionally approved for (clinical) applications and how software is usually deployed and constantly updated [6].. While this problem might also exist with software that is already deployed in other ways in medicine and health care, ML tools take it to a new level. Here, updates might involve newly trained algorithms with a new data background which might have achieved different performance benchmarks. How should this agile updating be weighed against traditional and more accurate, but slower ways of approving tools for clinical application? In the US, the FDA already reacted by creating easier paths for approvement for this kind of software but the success of this pathway is still indeterminated.

ML tools are said to be able to increase efficiency in hospitals and prevent unnecessary hospital visits, thus reducing pressure on care workers and doctors, which is certainly a good thing [7]. However, it will be important to

take a holistic approach towards health care in the future. ML tools are too often seen as the holy grail to solve problems when in fact they are just tools that will not tackle structural problems without using them to do so. For example, in current health care systems reduced workload of care workers has the potential to lead to a reduction in the workforce because it is a way to save money. However, this would then not lead to an actual improvement for patients but only to a potential financial reward. This can be seen as a trade-off between monetary outcomes and spendings on the one and the patients experience and care on the other hand. While this is an issue that is already existing, ML tools bring another perspective to it since they have the potential to increase as well as heavily decrease the patients experience in hospitals.

Some scholars argue that while a mistake by a human practioner only affects a small amount (often only one) people, a mistake by an algorithm that is deployed on many hospitals will happen more often and thus affect more people [15]. This could be seen as a trade-off between the scale of deployment of a technology and the severity of mistakes. However, this argument can also be seen as flawed since although one practioner might not repeat a certain mistake, other practioners not involved in this situation might well do. Thus, the only argumentation here could be that the errors are not as systematically spread as with ML tools, although even that might be an overstatement.

Another Trade-off can be seen between empathy in human practitioners and a more standardized way of tackling tasks in ML tools [15]. What do we understand as good health care, only the right diagnosis or psychological wellbeing during treatment, etc. as well?

Health care systems around the world are more or less privatized, depending on the country. However, in the case of ML tools a lot of research and development is driven by big companies like Alphabet or IBM [15]. This makes sense since those componies are driving ML research in general but it poses the question whether we want to give such an important issue completely out of public and into private hands. While the privatization of health care was already posing problems before ML tools and they are in fact seen as a solution for the existing problems [15, 21] the questioning of privacy and trust is increased as well. Thus, this can be seen as a trade-off between the speed of development - arguably, big tech companies will be fast in bringing ML tools to the market - and privacy and trust issues.

5 Conclusions

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

6 Arthur's ideas

6.1 tradeoffs, or maybe introduction?

6.1.1 Trade-offs in ML

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. (citation required). Non-maleficience, trade-off [18].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 [16, 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" [13, 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" [13, 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." [18, 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. Cite [2]. 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].

6.2 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.

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

6.3 New problems

We do not, however, argue that ML does not introduce any new ethical problems, but that strictly fairness-related problems pre-exist the proposed algorithmic solutions. 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 [15, 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 [15, p. 4] (automation bias, [16, 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 [16, 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 [18, 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 [15, p. 4].

References

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

[2] Andrew Chester et al. "Balancing utility and fairness against privacy in medical data". In: *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. 2020, pp. 1226–1233.

[3] 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.

[4] Francois Dionne and Craig Mitton. *Health Care Trade-Offs: A Necessary Reality For Every Health System*. Mar. 20, 2018. URL: %5Curl%7Bhttps://www.healthaffairs.org/do/10.1377/forefront.20180316.120106%7D (visited on 02/26/2022).

[5] Robert Eiss. "Confusion over Europe's data-protection law is stalling scientific progress". In: *Nature* 584.7822 (2020), pp. 498–499.

[6] Jianxing He et al. "The practical implementation of artificial intelligence technologies in medicine". In: *Nature medicine* 25.1 (2019), pp. 30–36.

[7] Denis Horgan et al. "Artificial intelligence: power for civilisation–and for better healthcare". In: *Public health genomics* 22.5-6 (2019), pp. 145–161.

[8] 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).

[9] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. "Discrimination aware decision tree learning". In: *2010 IEEE International Conference on Data Mining*. IEEE. 2010, pp. 869–874.

[10] Christopher J Kelly et al. "Key challenges for delivering clinical impact with artificial intelligence". In: *BMC medicine* 17.1 (2019), pp. 1–9.

[11] Amirhossein Kiani et al. "Impact of a deep learning assistant on the histopathologic classification of liver cancer". In: *NPJ digital medicine* 3.1 (2020), pp. 1–8.

[12] John Launer. "Medicine and the art of trade-offs". In: *Postgraduate Medical Journal* 96.1139 (2020), pp. 575–576. ISSN: 0032-5473. DOI: 10.1136/postgradmedj-2020-138575. eprint: https://pmj.bmj.com/content/96/1139/575.full.pdf. URL: https://pmj.bmj.com/content/96/1139/575.

[13] 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.

[14] Aditya Krishna Menon and Robert C Williamson. "The cost of fairness in binary classification". In: *Conference on Fairness, Accountability and Transparency*. PMLR. 2018, pp. 107–118.

[15] 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.

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

[17] Pranav Rajpurkar et al. "AI in health and medicine". In: *Nature Medicine* (2022), pp. 1–8.

[18] 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.

[19] David F Steiner et al. "Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer". In: *The American journal of surgical pathology* 42.12 (2018), p. 1636.

[20] Reed T Sutton et al. "An overview of clinical decision support systems: benefits, risks, and strategies for success". In: *NPJ digital medicine* 3.1 (2020), pp. 1–10.

[21] Eric J Topol. "High-performance medicine: the convergence of human and artificial intelligence". In: *Nature medicine* 25.1 (2019), pp. 44–56.

[22] Stephen Toulmin. "How medicine saved the life of ethics". In: *Perspectives in biology and medicine* 25.4 (1982), pp. 736–750.

[23] Robert C Williamson. "The AI of Ethics". In: *Machines We Trust: Perspectives on Dependable AI*. Ed. by Marcello Pelillo and Teresa Scantamburlo. MIT Press, 2021. Chap. 9, pp. 139–160.

[24] Indre Zliobaite. "On the relation between accuracy and fairness in binary classification". In: *arXiv preprint arXiv:1505.05723* (2015).