Ethical trade-offs in medical Machine Learning technologies

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

March 5, 2022

Abstract

To change!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 [37, 44]. 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 different ways [38], 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) [34].

CDSS have been used since the 1980s with growing success [42]. 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 [42].

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 [18, 45]. 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 [45]. 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" [45].

Furthermore, medicine poses some inevitable moral dilemmas that can not be ignored, about which a substantial moral literature has been developed. The high stakes found in some decisions in medicine might induce less care in implementing ideas about 'fair distribution' and less readiness to sacrifice performance for fairness. In health care settings, we are often interested in the prevention of harm rather than the allocation of goods (such as in job placements, college admissions, and other areas where fair-ML is being applied). This means that solutions such as randomization of a group's predictions to ensure equal harms will be considered for example unacceptable. A final point of interest is the fact that in medicine, sensitive attributes such as gender and race might be true predictors and avoiding group differences would harm everyone [32].

In our investigations, we draw on literature from the fields of fair-ML, economics, medicine, biomedical ethics, and philosophy. In Chapter 2, we examine economists' work on the notion of trade-offs to derive a meaningful formalization of them. We then examine

trade-offs that are frequently discussed in the ML literature. The analysis of the medical literature found in Chapter 3 helps us identify the trade-offs (often described as moral dilemmas) inherent to the field of health care. We furthermore examine fairness problems in medical practice, and discuss the usefulness of principled approaches found in the bioethical literature in enlightening the ethical discussion about CDSS. In Chapter 4, we relate fairness problems raised by the fair-ML community to ethical questions raised in the medical field. In particular, we argue that many problems are inherent to medical practice and not introduced by the application of ML. We then examine what ML can positively contribute with respect to ethical issues in medicine. Finally, in Chapter 5 we show how the application of ML to health care introduces new trade-offs and ethical questions. We analyze the supporting literature critically, by discussing in which measure those questions can be found in ethical discussion about other technologies.

2 Introduction to trade-offs

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 ("the central rationale for many policies" [18, p. 77]). 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 [28]. Similarly, policies dealing with large-scale stochastic problems (vaccination, traffic security, nuclear deterrence, criminal justice) always entail harms and benefits [18], and hence trade-offs.

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 optimality

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 as a maximization in the following way:

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

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 maximize 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

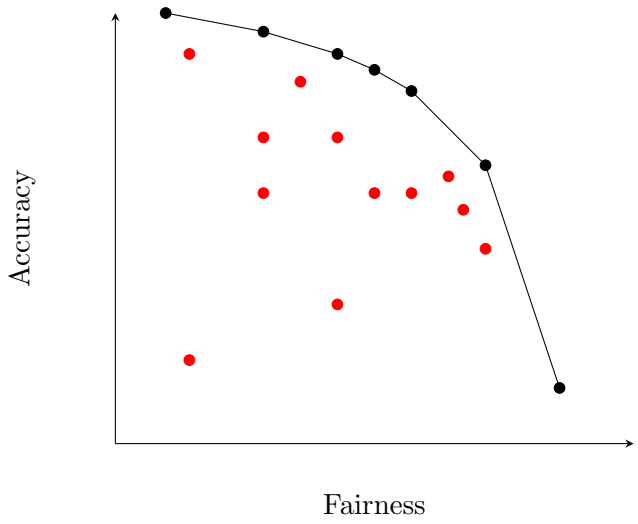

Figure 1: placeholder

solution $x' \in X$ if the following both hold:

1. $\forall\, i \in 1, 2, \ldots, k : u_i(x) \geq u_i(x')$
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.

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 [47].

By applying the concept of Pareto optimality one could say that any decision that doesn't involve a trade-off of some sort would be trivial to resolve, because it would have a unique, unconstrained maximum. Of course it would be desirable to avoid many of trade-offs in the sense of maximizing all objectives simultaneously, but that maximizing solution might not be in the feasible set.

...

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
- other MCDM approaches

Also write sentence about algorithms to calculate front in high dimensions

2.2 Trade-offs in fair Machine Learning

When designing any technology [1] there are many trade-offs inherent in the process. Of course machine learning systems and algorithms are no exception.

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

...

[7]

Here we are going to characterize three main fairness related trade-offs we identified in machine learning systems:

Choice of fairness measure Often we will quantify fairness subject to a selected fairness measure. Some often used examples include equalized odds, statistical parity and predictive parity [17]. The factors to take into account when deciding on what metric of fairness to use for a specific ML system are multiple. The different metrics each formalize different notions of morality [5]. First we need to decide what moral principles we want to follow, i.e.what we intend by equal or just treatment. But the choice of measure entails a trade-off already [8] [27] [3], as some notions of fairness are mutually contradictory and cannot be satisfied at the same time. In turn we have to choose between several contradictory measures, given that they are feasible for the ML problem at hand. All the metrics mentioned above are group fairness measures. In group fairness we try to protect so-called sensitive attributes, in the sense of treating different groups (that are separated by the sensitive attributes) equally. But how do we choose the sensitive features or groups we want to protect? We cannot take the sensitive features for granted, instead the separation into groups is also always a choice that humans have to make.

One apparent solution to the problem of choosing a group fairness measure could be to instead consider so called individual fairness. Here we consider equal treatment of similar individuals, independent of any group affiliation. By only considering individuals instead of groups we seemingly sidestep the problem of choosing the sensitive features we want to protect and which fairness measure to use. Choosing an appropriate measure for the similarity of individuals poses difficulties, as there are many feasible measures and we cannot restrict the space of metrics by imposing additional (moral) constraints on the choice of measure. But certain similarity measures do correspond to certain moral notions of equal treatment, e.g.statistical parity [14].

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 discrimination based on sensitive features.

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.[24] and [52] avoiding discrimination (or achieving a certain level of fairness) often directly results in the loss of prediction accuracy.

Furthermore, [33] showcases that this trade-off 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. They show that the accuracy and fairness trade-off 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 fully 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 examine the Pareto front of the trade-off. This is characterized by [29] and [49], where the trade-off between accuracy and fairness is given as a Pareto front for different measures of fairness. This allows for examination and comparison of the nature of the trade-off for different problems and measures.

Accuracy vs. Privacy Privacy, just like fairness, is another information-based harm [46], albeit with slightly different characteristics.

Van den Hoven [46] characterizes four main moral reasons to protect privacy and personal data:

elaborate this...

For example in a medical application a patient would not object to his data being used for their own treatment, but might be opposed to be disadvantaged in another context (e.g.the workplace) based on that same medical data.

...

In the simplest approach to the problem, to assure a certain degree of privacy we have to discard/randomize some of the data, which in turn will lead to worse accuracy. There have also been proposed multiple more refined approaches to mitigate this loss of accuracy (many approaches for privacy preserving data mining/machine learning e.g.[13] using differential privacy).

But while those are able to preserve accuracy to some degree under certain privacy guarantees, they aren't always applicable and don't necessarily resolve all of the harms mentioned. This can be partially explained by the different meanings of "privacy" in dif-

ferent fields. Differential privacy also might additionally amplify the unfairness of a given model [2], which is especially problematic when we try to balance all three axes of conflict (accuracy, privacy and fairness) in our ML system [7].

3 Trade-offs in medicine

Let us now turn to trade-offs that can be observed in the practice of medicine. In this chapter, we concentrate on trade-offs found before the application of Machine Learning.

3.1 Trade-offs in traditional medicine

As already hinted to earlier, trade-offs 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 different possible treatment options for one specific patient. 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 [50]. Furthermore, the simple concept of trade-offs in cases where sacred values (such as human lives) clash with secular values (such as money) is often morally disturbing and avoided [43]. 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 [51]. Hence, a more accepted term here is ethical (or moral) dilemma, which is a problem that arises when opposing values or principles co-occur [40]. 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. 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[35]). In this case the physician faces multiple trade-offs: 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). Far from purely qualitative reasoning, a step in the quantitative dimension of trade-offs is shown for example by evidence-based medicine [28], 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 [28]. In fact, one can go as far as to "conceptualize medicine itself as the art of managing trade-offs" [28]. 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.

Health care itself, as part of governance, is ridden with trade-offs. Health care systems 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) [11]. Empirical research suggests that the phrasing of such decisions has a big

impact in the public's perception of the problem: "Hospital administrators wrestling with tragic trade-offs can find themselves in the dock as soon as critics wonder who set the budget constraint that made it possible to save only one child" [43, p. 323]. This might explain the resistance to trade-offs talks in health and preference for moral dilemmas talk mentioned above, despite the arguably higher practical utility of trade-off concepts.

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 [15].

A more hidden trade-off, masked by claims of complete objectivity of science, is the one between methodological criteria in clinical research. With the modern focus on evidence-based medicine and randomized controlled trials and consequent diminution in value of cohort studies, case-control studies, expert opinion, and case studies, active (interventional) studies have been elevated to the golden standard of medical research. However, in doing so, the whole focus is placed on the methodological criteria of generality and precision, disregarding criteria such as realism, coherence, explanatory power, and others. This hides the underlying trade-off between methodological criteria, giving an absolute choice where case-by-case considerations of medical focus and contextual values are required to set priorities [21].

3.2 Fairness in traditional medicine

We can identify some key fairness issues in the medical literature. 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)" [6, p. 2]. Furthermore, gender bias is a recognized factor in health care. It is observed for example in the uneven composition of clinical trials samples, concentration on male-typical risk factors in studies, and in the different seriousness with which men's and women's complaints are received by medical doctors [39].

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 [40], [34], [36]. 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 [40].

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[45] 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. Physicians typically exert their clinical judgment only after collecting a precise case history, instead of following general theoretical considerations early on [45].

4 Combining machine learning and medicine

Having identified a number of trade-offs entailed in the practice of medicine, in particular ethical ones, we now turn to problems that were identified by the fair-ML community when developing support systems. We first provide a short summary, and then argue that most of the emerging concerns are not fundamentally new, but related to older ethical problems intrinsic to (medical) decision problems. After arguing that the performance and fairness of CDSS should be analyzed not absolutely, but relatively to what human practitioners currently do, we evaluate what contributions and improvements (as it relates to fairness) ML could bring to medicine.

4.1 Fairness problems at the intersection of Machine Learning and medicine

Ethical questions about algorithmic unfairness are a type of normative concerns [34]. The general concern is that CDSS trained on unbalanced or biased datasets might pick up the wrong patterns and exacerbate existing inequalities in health by overfitting on advantaged groups [6, 34]. The problem is often identified in the data used for training, which might contain label prejudice (a kind of negative legacy), variability amongst clinicians and institutions, and evolving clinical knowledge [6]. The data at deployment time is also identified as a source of problems: Population shift makes the developed model not adapted to the current population, and the usual lack of ground truth labels at test time makes evaluation difficult [6]. Ethical discussion emerges in classification problems, when the operating point of the algorithm is chosen [16]. This corresponds to the trade-off between false positives and false negatives, a well-known question in medicine (for example in breast cancer screening [16]). While bias is an ethically neutral term indicating statistical imbalances, unfairness is the judgement of bias as incompatible with moral principles [16]. Examples of biases of interest are sampling bias, unforeseen correlations, true systemic bias with biological causes, and batch effects [16], as well as negative legacy and labeling prejudice [6] So it must be recognized that "algorithmic development is never an entirely objective, value-free endeavour: it will be influenced by a host of social and institutional norms, practices and attitudes that could well build bias into design." [51, p. 673]. Fairness in ML is usually defined in terms of groups, quantified by group fairness measures. The consensus is that fairness through unawareness is not the right solution, because of sensible attribute leakage and the true effects of latent biological factors in many diseases [6].

4.2 Old and new problems

The practical problem of applying ML to health care decision 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 [16], and in our case must draw on medical ethics literature [34]. In particular, "considerations internal to medical science" and "contextual values" must help inform the decision on which methodological criteria to focus [21, p. 252], and whether the available data and ML technology allow such a focus. Hence, we suggest abstracting from ML itself and regard it as a general technology used to solve a pre-existing problem. We thus take inspiration from philosophy of technology, recognizing the interaction between our behavior and the technology we use [40]. Our reason to, on a first analysis, disregard the specifics of ML is that in many cases, ML often simply specifies existing trade-offs and makes them unavoidable.

As a concrete example, take the problem of unbalanced training data causing the ML 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 [51]. 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 partly derived from observations from medical practitioners and/or statistical studies. Biased data informing those studies will bias the observed results [6]. 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 [32]. 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 [6] or by driving the collection of more diversified data [51]. 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 [32] (although traditional statistics can help, see e. g. importance sampling).

Technical bias, emerging from new knowledge or data not being integrated into the system, is arguably similar to emergent bias that is typical of any decision system. Human deciders are for example subject to the 'availability heuristic', and are often mandated continuing education to keep their knowledge up to date, to avoid emergent bias [51]. Similar mandated technical solutions (software upgrades, maintenance procedures) are in principle possible fixes for the problem [51].

The issue of reducing an individual to a group identity already exists in statistics, and

arises in classical clinical practice as well. For example, there is evidence for the strong moral resistance to the use of statistics (such as generalization-allowing base rates) in sensitive situations [43]. Despite claims about the possibility of 'personalized medicine' thanks to ML methods, individuals are still reduced to their features [4, 40]. Although it is true that personalization might be an illusion, ML allows for more precise groupings. This might be enough to counter objections about generalization, since "[w]hat appear to be criticisms of generalizations in general(!), may in fact boil down to criticisms of *insufficiently precise* means of generalization." [4, p. 5].

Similarly, the problem of data privacy pre-exists the entry of ML in the medical field [40]. Privilege bias (models being developed for diseases that disproportionately affect a certain group) [36] is a problem that exists in classical statistical studies as well [23]. Cognitive biases of all sorts (availability heuristic, anchoring effects, framing effects, tendency to see false correlations, wrong probabilistic reasoning especially with small probabilities) have been shown in decisions taken by humans [51]. In particular, availability bias and anchoring effect have been observed in medical diagnosis, with increased effects for more expert doctors who rely more on non-analytical reasoning [31]. 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. 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) [32, 36]. This however requires deciding on a fairness measure to enforce (and how much to weigh it against performance), that will in most cases be in contrast with other fairness definitions [51].

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' and 'equity'. 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. Fourthly, "it can significantly reduce one of two potential sources of bias and discrimination [...] *intrinsic* bias" [51, p. 672], by removing the influence of (unknown) prejudice and emotions from decisions (although they might still be present in the training data). Knowing that those goals influence the results [40] (for example, pure efficiency potentially leads to the propagation of health inequities [36]), the importance of each objective has to be decided upon (and with it, the 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 [50]. 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 [51]. Let us reiterate

that, in many cases, ML technologies would simply assist human decision-makers in existing tasks [34]. This is partly due to laws that make fully automated decisions impossible, such as the European Union's GDPR [51]. It must be noted that, while potentially being a solution to legal headaches, replacement does not remove the responsibility from the development chain of CDSS, since such systems might lead to overreliance on tools and deskilling of practitioners [34]. The analysis of the impact of ML technologies must be relative to the current human decision-making (probably taking a utilitarian point of view, as is typical for stochastic problems; see [18]), and not an absolute decision about whether they act 'perfectly fairly' or not (which is mathematically limited in every decision system anyway [51]). A partial reason for public distrust of algorithmic solutions might be a wrong image of doctors as invulnerable and perfect figures, partly protected by physicians in trying to keep a "symbolic facade of professional competence" (while privately recognizing the risks and errors of their practice) [48]. We argue that attention must be paid to always compare what is expected from algorithmic solutions and what is currently expected from the humans executing those tasks presently. In the case of explainability, an often-requested characteristic for ML systems in health, double standards can be shown between what is expected from ML and what practitioners currently do [51]. One can for example point out that "the human brain, too, is largely a black box" [51, p. 666], and that explanations for decisions are often provided ex post, are influenced by emotions, and reflect mistaken rationalizations [51]. A comparison to current unaided (by ML) practice is however rarely done, and is sometimes even impossible to do because of the impossibility and/or illegality of collecting statistics about human deciders [50]. Furthermore, it might be particularly hard to do fair comparisons in the public discussions, since generally "humans are judged by their intentions, while machines are judged by their outcomes" [20, p. 139]. So, as long as fairness problems shown in human clinical decision-making are seen as unintentional, it might be difficult to argue for an effective advantage of using algorithmic decision-making tools.

The visceral resistance to the use of any technological system that shows any behavior deemed as unjust might be stopping improvements in overall care, which can be considered problematic. Do we want, for example, to refute to apply any system that does not lead to equalized outcomes but only equalized benefit [36]? 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? Despite some resistance of doctors to evaluation (at least partially driven by fears of blame) [48], quantification of errors and fairness are needed to compare different solutions. Furthermore, the time efficiency and cost of the ML applications with respect to unaided clinicians must be taken into account. As discussed before, trade-offs are made about budget allocation and doctors' time prioritization, so reducing them might allow improvements in other sectors of medical by displacing the saved resources (which is a phrasing that might be more positively received than simply mentioning saved costs [43]). Especially considering the benefit potential of AI in medicine, exceptionalism for the application in this field is unjustified [16].

Since our view of the world is partial and hence stochastic, we have to accept that any decision is subject to uncertainty and hence probably imperfect. In particular, the inevitability of errors in the medical field (be them active or latent) is widely recognized by practitioners themselves [48]. The advantage of actively, empirically 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 and can be monitored at deployment time (failure auditing [6]). This would be particularly welcome in medicine, which suffers from a lacking culture of error culture as compared to industries such as aviation and nuclear energy [48]. While the common suspect for this lack is the "culture of blame" found in medicine, other factors can be found, such as the fear of external blame, the attempt to maintain a figure of competence, the normalization of errors, a revulsion to management, an individualistic culture, a skepticism about external non-expert observers, and collegiality [48]. 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 (instigating some kind of population-level behavioral change health [34]). For example, pushing to correct for historical bias by applying equal allocation principles [36] will give us more diverse data based on which to infer the causes of past differences, and potentially reduce performance gaps [6]. A possible solution is hence to develop the system that we deem the more appropriate with the current knowledge, accept the imperfection and improve them over time as they get results. In a high-stakes context such as medicine, it is likely that stronger performance-based auditing (evaluating the outcomes) will be needed, and not only accreditation-based auditing (judged by experts) [51]. ML systems are not tools that once applied will remain forever the same: They should be closely monitored and improved over time [36], potentially reducing the impact of dataset shift with the integration of new data. The requested standards, however, "*should* be applied consistently across the board, regardless of whether we are dealing with machines or humans" unless there is "some compelling political, economic or social justification to the contrary" [51, p. 678]. However, the dynamics of the entire ecosystem make it very difficult to predict its evolution. Furthermore, since the problems that are dealt with are stochastic in nature, the developed solutions will necessarily be stochastic and entail some kind of trade-off, typical of technological innovations [18]. Dealing with stochastic problems requires a weighting of benefits and risks by their probabilities [18], and statistical reasoning is exactly what ML is good at (and humans are not) [50]. Unfortunately, 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 [9].

5 New trade-offs

We have argued that most of the fairness trade-offs identified at the intersection of machine learning and medicine are not new, 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. This might involve trade-offs that uniquely emerge from the technology of medical ML [40], so let us zoom out of the fairness domain to see what is happening when ML and medicine are combined.

5.1 Trade-offs

ML applications in medicine are often discussed as a human vs. machine situation - where the medical ML system outperforms the human they should and in the near future will be substituted. However, this creates a binary decision that is hard to make, especially with ML systems which can involve a good amount of uncertainty. It also creates an environment were humans are competing with a machine for the prerogative of interpretation, which contains some understanding of ML systems as some kind of autonomous systems. This autonomy is rather imaginary since as of today, medical ML systems still need there decision making process to be started and evaluated by humans. Thus, we are left with a distinction between ML systems as a tool or a machine as it is described in [50] and as it is argued, this distinction is often made based on familiarity - new developments are machines and will only be called tools when they grow older and people get used to them.

Recently, different studies found that combining ML and human evaluation can achieve better results than either of the two on their own [26, 37, 41, 44]. This suggests that human practitioners should use medical ML systems to support their decisions and increase their efficiency rather than be replaced by them, which is in line with understanding medical ML systems as tools. 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 best accuracy, how much human involvement do we want or need? 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 judgment might need to be combined for the optimal solution. One of the studies mention above 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 [26]. Moreover, there is also evidence that AI might especially improve the performance of less experienced practitioners like those who are still in training while those who are already experienced would not profit as much [37]. This brings another level into the trade-off because experienced practitioners who already perform on a similar level as the AI the necessitation to use the technology might even be hindered by another step in their workflow. In fact, for CDSS without ML components it was already observed that more experienced doctors ignore its assistance more often without performing less good [42]. This trade-off is further complicated by the question of responsibility. Naturally, the more ML tool and physician interact the harder it gets to identify where potential mistakes come from and thus understand whether a mistake

by the practitioner or an error in the medical ML system are to blame [22]). This can be problematic since many applications of medical ML systems involve high-stakes scenarios where errors should be avoided at all cost or if they happen should be eradicated as fast as possible.

A clear responsibility framework is also important for fostering trust in the medical profession which leads us to the next trade-off. For a good relationship between patient and health care practitioner, trust is of the utmost importance [10]. For the application of ML in medicine and health we can identify a multi-faceted trade-off between trust in the system and accuracy that already starts with what we just mentioned, but continues far beyond that. It involves a trade-off that is standard to ML but grows to a great importance especially in applications like medical ML systems, namely the trade-off between explainability and accuracy [25, 44]. Most medical ML systems today perform worse as soon as some kind of interpretability framework is built in, leading to the question of how important explainability is for the application [30]. While it is not yet clear, how an explainability framework does influence the work of practitioners studies have shown that it would increase trust in the system, from the practitioners as well as from the patients side [12, 44]. This can also be related to another, rather philosophical trade-off: If a person is skeptical 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 technology more reasonable. If it is impossible to reach a conclusion, for example because there is not enough data, ML tools should be transparent about that and indicate that they cannot make a decision rather than making a bad informed decision. [22]

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 [36, p. 4].

How much do physicians need to understand the tools they are using? (Education of an AI-literate workforce [19]).

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 [22]. 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 practitioner 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 [34]. 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 practitioner might not repeat a certain mistake, other practitioners 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 [34]. What do we understand as good health care, only the right diagnosis or psychological well-being during treatment, etc. as well? ethical question: should we predict death? ([44] Table 3) [19] talk about triage by ML

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 [44]. 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 [19]. 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.

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 [34]. This makes sense since those companies 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 [34, 44] 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.

The development and deployment of ML tools in medicine and health care will and does already cost a lot of money [19]. 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 [44]. 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.

WTF? The savings would come from a combination of deployments: lower medical costs

and reduced losses from low productivity and sick day ([22] page 148)

Another trade-off exists between the way medical devices are traditionally approved for (clinical) applications and how software is usually deployed and constantly updated [19].. 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 improvement for this kind of software but the success of this pathway is still indeterminate. Today, regulation processes are often such that the model is locked in place before deployment. This makes it easier to regulate them but misses out on their potential to learn and increase functionality on the fly.

How do we study the clinical efficacy of ML tools? Is a randomized controlled trial ethical? Because with normal medical trials we do not have an alternative working treatment, we just compare it with nothing. Thus, we do not withhold something from patients. However, of ML tools (wrongfully) decide against treatment we actively withhold treatment from patients which might in extreme cases lead to their deaths. (Grote und Genin)

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. ([44] Increased Efficiencies) ([19] Transparency) Can digital health care be for everybody? what about people who do not have digital devices or don't want to use them?

But this also leads to the question whether these trade-offs are actually new or - as discussed for fairness trade-offs above - if they are just new editions of trade-offs humanity has seen before in either ML, medicine and health care or technology in general.

6 Conclusions

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

References

[1] Christopher Alexander. *Notes on the Synthesis of Form*. Vol. 5. Harvard University Press, 1964.

[2] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. "Differential privacy has disparate impact on model accuracy". In: *Advances in Neural Information Processing Systems* 32 (2019).

[3] Richard Berk et al. "Fairness in criminal justice risk assessments: The state of the art". In: *Sociological Methods & Research* 50.1 (2021), pp. 3–44.

[4] Reuben Binns. "Fairness in machine learning: Lessons from political philosophy". In: *Conference on Fairness, Accountability and Transparency*. PMLR. 2018, pp. 149–159.

[5] Reuben Binns. "On the apparent conflict between individual and group fairness". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 514–524.

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

[7] 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.

[8] Alexandra Chouldechova. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: *Big data* 5.2 (2017), pp. 153–163.

[9] 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.

[10] Chalmers C Clark. "Trust in medicine". In: *The Journal of medicine and philosophy* 27.1 (2002), pp. 11–29.

[11] 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).

[12] William K Diprose et al. "Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator". In: *Journal of the American Medical Informatics Association* 27.4 (2020), pp. 592–600.

[13] John C Duchi, Michael I Jordan, and Martin J Wainwright. "Privacy aware learning". In: *Journal of the ACM (JACM)* 61.6 (2014), pp. 1–57.

[14] Cynthia Dwork et al. "Fairness through awareness". In: *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012, pp. 214–226.

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

[16] Richard Ribón Fletcher, Audace Nakeshimana, and Olusubomi Olubeko. "Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health". In: *Frontiers in Artificial Intelligence* 3 (2021), p. 116.

[17] Pratyush Garg, John Villasenor, and Virginia Foggo. "Fairness metrics: A comparative analysis". In: *2020 IEEE International Conference on Big Data (Big Data)*. IEEE. 2020, pp. 3662–3666.

[18] Russell Hardin. "Ethics and stochastic processes". In: *Social Philosophy and Policy* 7.1 (1989), pp. 69–80.

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

[20] César A Hidalgo et al. "Moral functions". In: *How humans judge machines*. MIT Press, 2021, pp. 123–147.

[21] Vincent KY Ho. "Medicine, methodology, and values: trade-offs in clinical science and practice". In: *Perspectives in Biology and Medicine* 54.2 (2011), pp. 243–255.

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

[23] 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).

[24] 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.

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

[26] 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.

[27] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores". In: *arXiv preprint arXiv:1609.05807* (2016).

[28] 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.

[29] Suyun Liu and Luis Nunes Vicente. "Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach". In: *arXiv preprint arXiv:2008.01132* (2020).

[30] Yi Luo et al. "Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling". In: *BJR— Open* 1.1 (2019), p. 20190021.

[31] SÝlvia Mamede et al. “Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents”. In: *Jama* 304.11 (2010), pp. 1198–1203.

[32] 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.

[33] 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.

[34] 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.

[35] Bryan Oronsky et al. “Medical Machiavellianism: the tradeoff between benefit and harm with targeted chemotherapy”. In: *Oncotarget* 7.8 (2016), p. 9041.

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

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

[38] Will Rinehart and Allison Edwards. “Understanding job loss predictions from artificial intelligence”. In: *American Action Forum. Org.* 2019.

[39] M Teresa Ruiz and Lois M Verbrugge. “A two way view of gender bias in medicine.” In: *Journal of epidemiology and community health* 51.2 (1997), pp. 106–109.

[40] 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.

[41] 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.

[42] 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.

[43] Philip E Tetlock. “Thinking the unthinkable: Sacred values and taboo cognitions”. In: *Trends in cognitive sciences* 7.7 (2003), pp. 320–324.

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

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

[46] Jeroen Van Den Hoven. “Information technology, privacy, and the protection of personal data”. In: *Information technology and moral philosophy* 301 (2008).

[47] Irene Van Staveren. *The ethics of efficiency*. Tech. rep. 2007.

[48] Justin J Waring. "Beyond blame: cultural barriers to medical incident reporting". In: *Social science & medicine* 60.9 (2005), pp. 1927–1935.

[49] Susan Wei and Marc Niethammer. "The fairness-accuracy Pareto front". In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* (2020).

[50] 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.

[51] John Zerilli et al. "Transparency in algorithmic and human decision-making: is there a double standard?" In: *Philosophy & Technology* 32.4 (2019), pp. 661–683.

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