Guiding Practitioners for Managing Fair AI

Current Gaps, Tools and Future Action Needed

Savina Kim

Preface

The high social impact resulting from the use of AI in sensitive environments has raised major concerns over safety and fairness. As machine learning (ML)-enabled products and services are granted more autonomy in decision-making processes, they play a growing role in the distribution of critical resources and opportunities, ranging from credit allocation, parole, employment to even medical diagnosis. Thus, it is crucial to ensure these algorithms are free of discriminatory behavior and that fairness constraints are taken into consideration when designing and engineering these systems. While the interdisciplinary community has acknowledged this discourse and proposed several technical solutions, few incorporate pragmatic tools for industry application. Managers of real-world systems - eager to spark responsible AI use within their organization - receive limited support for how they may establish and manage these concerns. Against this background, areas of alignment and disconnect between the challenges faced by practitioners and the available methods in literature are first identified. Based on these findings, this paper proposes a fairness management framework known as the Fairness Health Check which can be used to evaluate ML systems with reference to a template workflow. It hopes to provide industry managers and practitioners with a starting point in managing fairness in AI and be illuminating to policy makers and regulators looking to address the needs consumers impacted downstream.

Contents

1 Introduction

It is now commonplace to see reports in popular press of the unfair systemic decisions made by now widely used ML systems - for example, facial recognition software being biased against darker-skinned individuals (Buolamwini, 2018), an automated hiring system recommending applicants of particular race, age or gender (Broek et al., 2019; Raghavan et al., 2020), a predictive policing algorithm claiming that "black people reoffend more," (Berk et al., 2017; Brennan & Oliver, 2013; Ensign et al., 2018), or the infamous case of Apple's 'sexist' credit card (Neil Vigdor, 2019). These technologies are progressively being employed for decision-making processes which can have a long-lasting impact on individual lives. Therefore, their propensity to unintentionally replicate, reinforce and amplify harmful existing human biases and exacerbate social inequities has incited major concerns over their fair and ethical use (Corbett-Davies et al., 2018; Hu & Chen, 2020; Liu et al., 2018).

In response, members of the research community have made great strides in identifying and quantifying (un)fairness using a variety of statistical definitions. Over 20 definitions have been proposed including efforts to differentiate and clarify their distinctions, such as the tutorial "21 Definitions of Fairness and their Politics" (Narayanan, 2018) and "Fairness Definitions Explained" (Verma & Rubin, 2018) as well as the introduction of toolkits encompassing fairness tests and bias mitigation solutions in order to operationalize fairness within AI systems. Well-known efforts include IBM with "AI Fairness 360," Microsoft with "Fairlearn," Google with its "What-If tool," and "Aequitas" from the University of Chicago to name a select few (Lee & Singh, 2020). These strides reflect the public's demand for more trust and transparency, calls by industry experts for more auditing tools to help data scientists identify sources of bias and unjust discrimination during the modeling process and the imminent tightening of regulation by governing institutions that is expected (E. O. of the President, 2016; Holstein et al., 2018; Raji et al., 2020a).

However, there is a lack of consensus around a singular definition for fairness with incompatible interpretations of what is considered "fair" and the fundamental limitations of simultaneously satisfying multiple definitions of fairness at once, shown by Chouldechova with their "impossibility results" (Chouldechova, 2017; Kleinberg et al., 2016). Broadly speaking, fairness is a desirable quality in society, however, a complicated task to achieve in practice due to its high context-dependency further convoluted by cultural and ideological differences, complex interdependencies between data features and competing objectives (e.g., profit, accuracy, speed) between the relevant stakeholders (e.g., industry experts, entrepreneurs, policymakers, regulators) during the decision-making process (Hutchinson & Mitchell, 2018; Mehrabi et al., 2019; Verma & Rubin, 2018).

Alongside the growing number of computational tools and techniques, public-private institutions and regulators have called for more societal accountability via the creation of high-level "AI principles." Over 84 initiatives have produced statements describing ethical standards for guiding the development, deployment and ultimately governance of these technologies; for example, that technologies should remain subject to human control and the creation or reinforcement of unfair bias should be avoided (Mittelstadt, 2019). Similarly, the General Data Protection Regulation (GDPR) and Data Protection Act 2018 require that organizations use personal data in a way that is 'fair,' however, do not define how the term should be interpreted in practice. This alludes to a major flaw in these efforts - these principles are too vague and provide little to no guidance on how to translate such principles into practice (Greene et al., 2019; Raji et al., 2020a; Whittlestone et al., 2019). Even for a skilled ML engineer, the multitude of considerations required to ascertain fairness accountability and overcome related challenges can be overwhelming. Therefore, in the absence of universal standards and formalized guidance, the desired alignment of AI principles with responsible model deployment is understandably lacking. This highlights an urgent need for internal processes and tools to support practitioners in developing fairer systems from the ground-up.

This paper asks the following question, "How do we encourage the application of existing fairness techniques by practitioners?" The proposed Fairness Health Check (FHC) is designed to enable organizations with a practical solution for managing AI as part of their social and ethical responsibility. Section 2 begins by introducing the practical considerations needed to assess fairness in ML pipelines and the related challenges hindering auditing approaches from being widely adopted yet. Section 3 introduces the FHC, a template which provides the basis for fairness assessment and helps facilitate more actionable analysis for practitioners and non-experts. Finally, the paper concludes with a discussion on how this self-regulation approach can be applied in practice, its implications on wider organizational and regulatory change as well as outstanding questions as potential avenues for future work with the goal of improving accessibility to this domain for both expert and non-expert users going forward.

2 Auditing AI

Faced with complex choices and resulting trade-offs, companies have and may continue to respond by deliberately remaining vague about their model details or deferring this task to a future date when industry standards are mandated, or worse, entirely abandoning the fairness effort altogether. However, while there is no universal means to ensuring fairness yet, there is optimism in the growing community working on operationalizing fairness at scale. Recent efforts have called for internal audits, similar to quality assurance measures, which can be used to enrich, validate and guide (fairness) risk analysis during model development (Morley et al.,

2021). An auditing process can ultimately boost confidence and guarantee the algorithm's trustworthiness by determining if it meets the proper regulatory, governance and ethical requirements alongside encouraging the adoption of mitigating measures (Sandvig et al., 2014). In other words, to evaluate whether a candidate automated decision-making system, once applied to real-world data, will operate within the expected behavioral standards and with the appropriate degree of transparency.

A few immediate considerations are first noted. Audit results are sometimes approached with skepticism because they are reliant on human judgement and vulnerable to interpretation. This implies that validating its result requires first establishing an audit integrity procedure which must adhere to a fixed, vetted methodology, similar to those observed in tax compliance auditing (Raji et al., 2020b). If this is ensured, AI auditing can become one of the most effective strategies to encourage (or mandate) companies to respect and incorporate ethical and responsible AI principles within their established ML workflow. Another concern revolves around the question of internal versus external auditing procedures, where the latter implies companies are accountable to a third party (Raji & Buolamwini, 2019; Sandvig et al., 2014). However, external auditing is fundamentally limited by the lack of access to the internal, proprietary details of a model such as the input data, hyperparameter choices, intermediate models and other design nuances related to the end users or consumers impacted downstream. Therefore, internal audit methodologies are advantaged by having direct access to the ML system and thereby able to incorporate information typically inaccessible to external evaluators which can help further reveal unknown risks.

The noted lack of reliable, fixed standards for auditing procedures poses as the first major roadblock hindering the advancement of regulatory efforts today. This is where improving the identification, documentation and tracking of the systemic and cultural requirements needed for fairness can serve as a remediation. Therefore, the remaining paper focuses on the advancement of audit integrity, in particular a fit for purpose fairness framework using a model template to offer a guided approach to AI fairness management. The following section introduces the framework which drew inspiration from annual vehicle inspection checks (also known as MOTs), which when applied to the needs for AI fairness safety, encourage closer inspection of factors critical to the overall future "safety" of the model. These can then be rated according to the level of urgency needed.

3 Fairness Health Check

The following provides an overview of the methodology used to build the Fairness Health Check (FHC) framework. The FHC represents a synthetic indicator obtained when individual indicators (fairness metrics) are assessed individually and encompass three critical components: a transparent methodology, sound statistical principles and coherence to relevant fairness theory. FHC helps draw attention to areas in need of further investigation by visually framing a selection of fairness metrics against the current state of the model. Fairness can be interpreted at three distinct but related levels: product, policy and implementation (Bakalar et al., 2021). FHC focuses on fairness at the implementation level which looks at the empirical performance of the system, relating to questions such as, "Is the predictive model achieving the desired tradeoff between different types of errors for all subpopulations, both unprivileged and privileged groups?" Assessing disparities between different protected subgroups is paramount for monitoring and preventing the potential harms of automated decision-making systems. This is accomplished by evaluating the difference in decision outcomes received by each group, either marginally or conditional on some ground truth. Although what amount of unfairness is considered 'acceptable' depends on the appropriate ethical, regulatory and legal context and may differ per application; in any case, it must first be measured.

3.1 Desired Properties

In search of an adequate framework for fairness the following goals are set. In order to encourage adoption and ensure utility, the desired properties are the following:

1. **Simple and easy-to-use** for any organization, thereby not requiring extensive resource allocation or expertise for implementation
2. **Clear, actionable and meaningful** items which provide a snapshot summary on the model's current state
3. **Consistent design** which can be iterated multiple times during model development and deployment allows for easy monitoring of trends over time therefore facilitating inter-model comparison
4. **Flexible yet specific** enough to enable managers and practitioners to challenge and evaluate complex organizational decisions needed to satisfy fairness concerns

3.2 Dimensions of Fairness

A selection of fairness measures was included as indicators after surveying the existing academic literature, where over 26 definitions for fairness have been proposed across various

sources (Corbett-Davies et al., 2018; Hardt et al., 2016; Mehrabi et al., 2019; Verma & Rubin, 2018). While these definitions are derived for multiple purposes, defined by different parameters, coverage and sociological requirements, statistical observation-based fairness metrics are focused on due to their reliance on model outcomes only and their relative ease for comparison purposes. This methodology hopes to encourage further investigation into the modeling pipeline which can then be revised utilizing the appropriate pre-, in- and post-processing techniques for targeted improvement (Bellamy et al., 2019; Raghavan et al., 2020).

Fairness definitions generally fall under two categories: group fairness and individual fairness. The former looks to protect vulnerable groups based on the distribution of classifications and errors and is therefore typically expressed as a "balance" or approximate equality of a select statistical measure between groups (Green & Hu, 2018). The latter focuses on ensuring two similar individuals receive similar outcomes (Dwork et al., 2011; Kusner et al., 2017). Whether these two notions are mutually compatible remains ambiguous (Binns, 2019). The following details select group fairness definitions and their corresponding explanations, which are sourced from the distribution of members across the target classes (see Table 1 with a lay example related to the credit application context).

Metric Name	Formula	Credit Example
Positive Predictive Value (PPV, Precision)	TP / (TP + FP)	Probability of customer with a good predicted repayment behavior to actually be good
Negative Predictive Value (NPV)	TN / (TN+FN)	Probability of customer with a bad predicted repayment behavior to actually be bad
False Discovery Rate (FDR)	FP / (TP+FP)	Probability of customer with a good predicted repayment behavior to actually be bad
False Omission Rate (FOR)	FN / (TN+FN)	Probability of customer with a bad predicted repayment behavior to actually be good
True Positive Rate (TPR, or Sensitivity)	TP / (TP + FN)	Probability of customer with a good repayment behavior to be correctly assigned
True Negative Rate (TNR or Specificity)	TN / (FP + TN)	Probability of customer with bad predicted repayment behavior to be correctly assigned
False Positive Rate (FPR)	FP / (FP + TN) = 1 – TNR	Probability of customer with bad repayment behavior to be incorrectly assigned to good class (falsely accepting a negative case)
False Negative Rate (FNR)	FN / (FN + TP) = 1 – TPR	Probability of customer with good repayment behavior to be incorrectly assigned to bad class (falsely rejecting a positive case)

Table 1. Statistical measures of fairness.

A selection of the most commonly cited group fairness definitions is detailed below:

- **Statistical Parity (a.k.a. Demographic Parity or Group Fairness)** (Dwork et al., 2011) is satisfied when members of both vulnerable and not vulnerable groups have equal probability of being assigned to the positive class; the outcome is independent of the protected feature.

- **Conditional Statistical Parity (a.k.a. Conditional Demographic Parity)** (Corbett-Davies et al., 2017) extends the previous definition by allowing a set of valid attributes to affect the predicted outcome; it is satisfied when members in both groups have equal probability of being assigned to the positive class, after controlling for a set of justified variables (e.g., credit history, requested credit amount or employment status).
- **Predictive Parity (a.k.a. Outcome Test)** (Chouldechova, 2016) is satisfied when members of both groups have equal PPV.
- **False Positive Error Rate Balance (a.k.a. Predictive Equality)** (Chouldechova, 2016; Corbett-Davies et al., 2017) is satisfied when members of both groups have equal FPR.
- **False Negative Error Rate Balance (a.k.a. Equal Opportunity)** (Chouldechova, 2016; Hardt et al., 2016) is satisfied when members of both groups have equal FNR.
- **Equalized Odds** (Hardt et al., 2016) is satisfied when members of both groups have equal TPR and FPR.
- **Disparate Mistreatment** (Zafar et al., 2016) is satisfied when members of both groups have equality of error probabilities.

Many argue that fairness cannot simply be captured by notions such as statistical parity and no mathematical formula alone can encompass the vastness of the issue (Green & Hu, 2018; O'Neil & Gunn, 2020). In other words, no single metric is a silver bullet; however, while mathematical formulations alone cannot guarantee just outcomes and one must also look "beyond the numbers," measurement is a necessary first step for progress. The use of statistical means helps practitioners understand what is occurring on the ground so that in conjunction they can think critically about the implications of any difference observed and further improvements can be made.

3.3 Protected Attributes

Before discussing the key components of the FHC, a brief mention should first be given to the notion of protected or sensitive variables (used interchangeably). Common examples include ethnicity, gender, race, religion, sexual orientation, disability and age which represent divisions in the data which are socio-culturally precarious when applying ML-based models. Disparities of social outcomes can often be attributed to populations identified as either underrepresented or vulnerable. The former being individuals within ethnic minority or gender groups which have been historically subject to discrimination and therefore are underrepresented in current society such as in workforce participation while the latter represent those lacking the social capital to adequately represent themselves making them vulnerable such as students, incarcerated individuals or those who are economically deprived and alienated (Kuhlman et al., 2020; Shivayogi, 2013). Many of these variables are explicitly defined

as "sensitive" by specific legal frameworks (Berk, 2019; Marjanovic et al., 2018; Sokolovska & Kocarev, 2018; Yeung, 2018) and are prohibited from being collected or utilized as features in the decision-making process in certain contexts. Anti-discrimination laws, also referred to as fair lending laws (e.g., Fair Credit Reporting Act (FCRA), Equal Credit Opportunity Act (ECOA) and the Fair Housing Act (FHA)) attempt to maintain a vision for fairness in the US with similar consumer protection directives implemented in the EU with Articles 12 and 13 EC as well as the Basel Banking Agreement. Additionally, the General Data Protection Regulation (GDPR) makes up the most significant legal instrument governing consumer credit scoring regulation with regards to data privacy as well as maintaining certain provisions targeting automated decision-making. In order to quantify disparity, the target (un)privileged groups, representing a subpopulation often defined by one or more of the sensitive variables, must first be selected where each template corresponds to one protected attribute of interest. This enables a direct comparison between the unprivileged (UG) and privileged (PG) groups.

3.4 Computation

The FHC consists of an index of fairness metrics designed to consider the position of the unprivileged group compared to that of the privileged group; for example, women to men or minority to non-minority group. This is computed as a ratio γ and constructed as follows:

$$\gamma = \frac{X_i^{UG}}{X_i^{PG}}$$

for the fairness metric X and model i. Since the unprivileged group's performance is always in the numerator, the ratio accumulates inequality in one-direction allowing relative compensation to the group. If the ratio is equal to 1, there is equality between the unprivileged and privileged groups whereas if the ratio < 1, the unprivileged group is discriminated against (e.g., women are disadvantaged compared to men) and if ratio > 1 the privileged group is discriminated against (e.g., men are disadvantaged compared to women), as visualized in Figure 1.

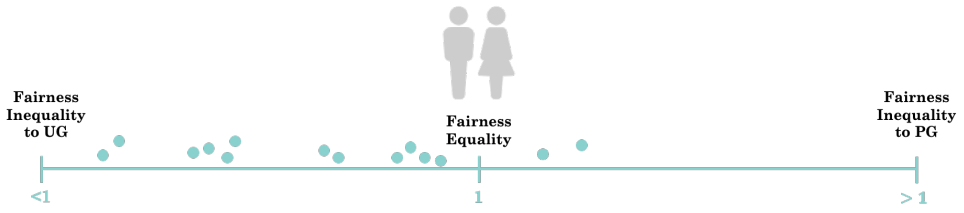

Figure 1. Fairness inequality ratio.

3.5 Thresholding

Taking inspiration from a MOT design (an annual test required for vehicle safety), a similar design is utilized but in this case of fairness safety. The FHC includes a key for fairness failure with three colors representing the level of urgency needed for each condition. Based on high, medium or low index scores, green indicates 'Acceptable' meaning checked and OK at this time, amber indicates 'Advisory' meaning may require future attention, and red indicates 'Urgent' meaning requires immediate attention respectively. By referencing the four-fifths rule, a popular measure used to identify adverse impact (Uniform Guidelines on Employee Selection Procedures), the green category encompasses any absolute difference less than 20% between the UG and PG. In other words, the index score of the UG should be within at least 80% of the respective index score of the PG. Amber is selected when an absolute difference in the range of 50-79% is found and the remaining cases until perfect inequality at 0 are designated as red, indicating urgent attention is required (Table 2). For example, dividing the true positive rate for females (72%) by the higher (95%) of males, results in 76% (72/95) which is lower than the legal minimum 80%. Therefore, the yellow category "Advisory" is marked for further investigation by the manager and relevant team members.

Key	Criteria (Absolute Difference)
Green	80 - 100% (fairness equality)
Amber	50 – 79%
Red	0 - 49% (fairness inequality)

Table 2. Criteria used for selecting level of urgency.

3.6 Template Design

Compiling all the fairness metrics and computations listed above, a comprehensive template was designed to help practitioners assess the fairness status of their current model. Having a standardized framework is useful for measuring fairness inequality over time which allows managers to track improvement trajectory as well as changes across iterations in a model's design pre-deployment. The complete FHC template can be found below:

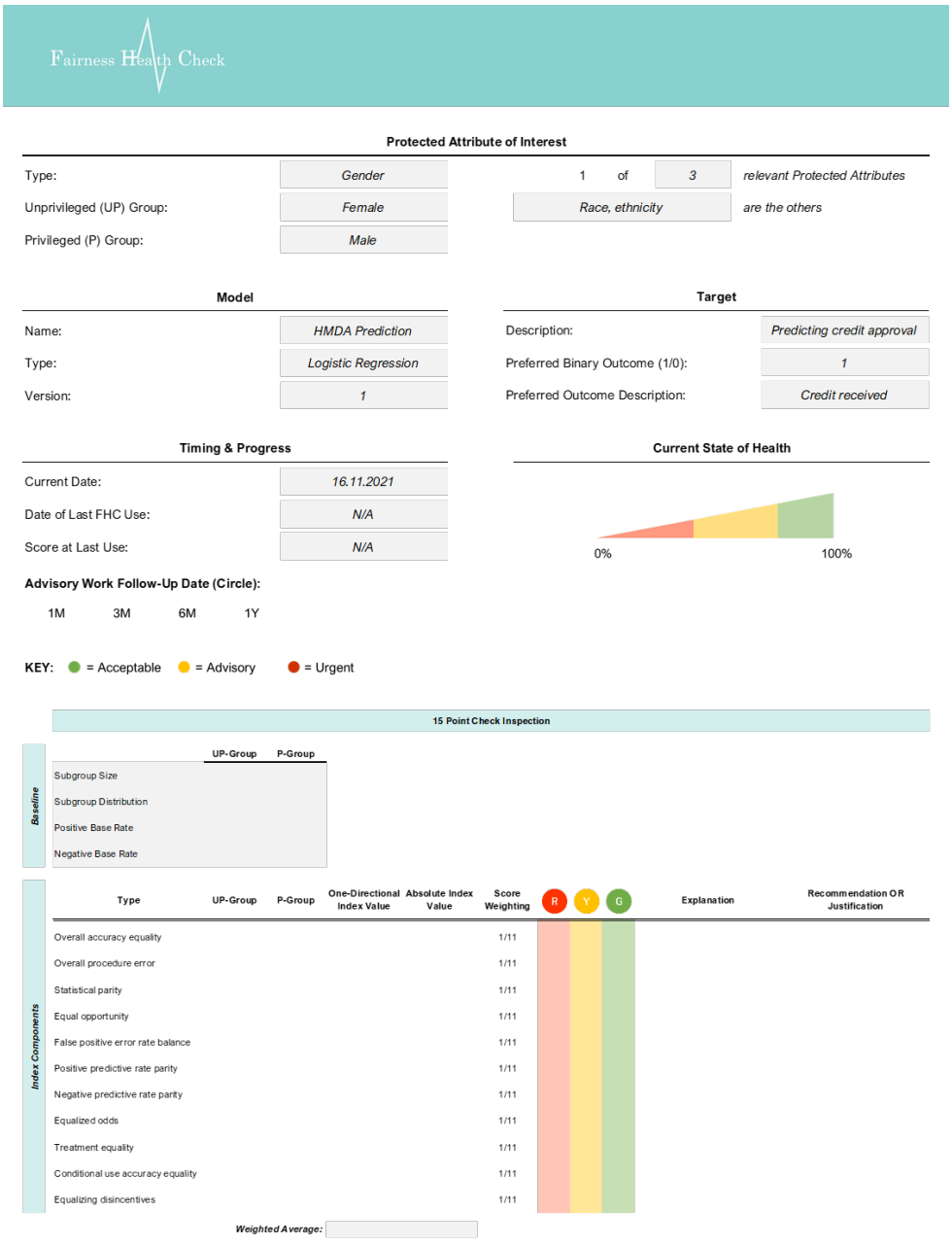

Figure 2. Fairness metric template.

This template can be utilized for multiple binary comparisons between the unprivileged and privileged subgroups relevant to the dataset. This is tracked in the upper half of the FHC where the user is asked to provide a description of the target protected attribute as well as other relevant attributes in the dataset for which this procedure can be replicated. In order to encourage iterative use of this template and track changes over time to drive improvement, the "Timing and Progress" section enforces the user to include previous scores as well as the expected date of the next follow-up fairness check. Furthermore, by framing the requirements in

a checklist-like design, FHC can help practitioners have a more comprehensive view of the important failures, edge cases and questions outstanding. In order to reduce the risk of FHC becoming an ineffective box-ticking activity, it is designed to ask practitioners to provide a descriptive explanation (lexical translation) for each calculated outcome as well as a written description of the type of improvement needed as a result or justification for the disparity. By guiding practitioners to describe their assessment of fairness risk, the FHC acts as a safety tool to help organizations avoid higher-level model hazards and thereby real-world failures, such as the 'sexist' Apple credit card.

To summarize, the FHC provides a user-friendly visualization that highlights specific areas in need of further investigation originating from a failed fairness metric. It allows the user to identify the range, or maximum possible bounds, of disparity that is learned from the data, which can then be targeted for improvement or (at the very least) require organizations to provide justification. It should be emphasized that this is, however, not a call to simply take incompatible notions of fairness and just "average across them." This would produce irresponsible results; instead, the FHC should be used to help answer questions such as, "Is this model fair to subgroup X?" and if so, according to which normative principle? By balancing across success rates, error rates and proportionality with pre-existing prevalence and surveying the overall effects on distinct subgroups affected by the model, this serves as the first step in helping practitioners calculate the social benefits and costs of their model and ensuring a consistent and holistic response to them, as opposed to harmfully ignoring them.

4 Discussion

There are a few limitations of the FHC that are worth noting, as they relate to the existing open challenges faced by researchers and practitioners today. First and foremost, the use of checklists during system development comes with several cautions, particularly the risk of blind application where the broader, systemic or societal contexts are neglected. In other words, imposing statistical definitions of fairness on ML models without acknowledging the wider ecosystem in which they are embedded in can quickly backfire and undermine fairness as opposed to fostering it. This framework does not imply that companies should move directly from its outputs to conclusions on their fairness policy and field of action; instead, this framework should be used as an information tool and one aspect of their broader holistic strategy. For example, a perverse effect of this framework would be if companies were to instrumentalize this as a tool for legitimizing company policy by restricting their response to fairness concerns in ML models to what is simply measured by the framework. In doing so, companies may fail to consider other key aspects, such as overrepresentation of women users for certain financial products or existing inequalities in average loan sizes within certain ethnic

populations; these nuanced, interrelated considerations are equally critical to the overall fairness discussion.

However, with that said, it is helpful to have easily coherent metrics with well-defined target thresholds which the FHC provides, further enhanced by its transparent process. The proposed FHC framework helps practitioners organize their fairness assessment in modeling projects while aided by a visualization tool for guided exploration. In other words, it provides a procedural solution for meaningful model improvement which requires being able to clearly track changes over time or across models. This not only enables a "transparency trail" of documentation at each stage of a model's development cycle (Raji et al., 2020b) but can also serve as a tool for internal education and training on responsible AI and ethical awareness across multiple teams due to its usability for experts and non-experts. The following discusses the industry-wide implications of operationalizing fairness and outstanding gaps where further attention and research is still needed.

4.1 Open Communication in the Workplace

Ultimately, the quantification convention of different fairness notions provides a dual function: defining fairness for the specific decision-making context and influencing the field of action of a company. The FHC contributes to defining how fairness should be evaluated for managerial purposes, notably by highlighting key areas of fairness alignment and/or misalignment during various stages of the modelling pipeline and most importantly, encourages improved communication within organizations about the topic. A multi fairness metric-based design can be used to (1) determine whether the highlighted pain points originate from faulty modelling design or application which can be immediately remedied or are instances of more broader, systemic issues such as historical context and (2) provide evidence to effectively persuade team members that these are issues needing to be addressed. One of the goals of this framework is to oblige companies to communicate more transparently about their fairness gaps and thereby encourage them to act, a requirement which will eventually be mandated by more stern regulation in the future. Currently, regulators provide only general terms and conditions regarding fairness and responsible use of AI. Until more detailed guidelines and standards are outlined, the FHC presents an intermediary solution to prompt action across a range of users and applications, who are invited to accept this mantle of responsibility.

Moreover, major benefits can also be gained by the companies themselves who are looking to avoid being in tomorrow's headline accusing them of mass discrimination. By displaying their fairness concerns, goals and remedies, accusations of unfairness and disputes

over the nature of bias in predicted outcomes like those seen with COMPAS by investigative journal ProPublica can be avoided (ProPublica, 2016). In other words, it is preferable to have an open discussion with the public and moderate the justification of using a specific fairness metric as opposed to giving off the impression that the company simply ignored the issue altogether or did not consider such an effect in the first place. Accountability in this respect, as opposed to the specific methodology used, is central.

4.2. Benchmarking and Competition

The FHC also serves as a stepping stone for industry to determine where baselines should begin and where averages fall in hopes of starting a discussion around target measures and long-term ideals to strive for. For example, if most companies are disparately scattered across the board, this would indicate a lack of overall fairness norms. Alternatively, in the case most companies are within similar ranges, this would indicate a baseline already exists and the next step is improving overall averages past a specified "ideal" threshold. This would enable individual companies to recognize their relative standing within an industry-based or application-specific context - either being a champion of fair technologies or severely falling behind its competitors, which can have negative downstream effects on customer acquisition or retention. And similarly, companies and governments that become aware of, or are publicly known for, unfairness or discrimination in their decision outputs are more likely to carry out policies to reduce this disparity. Therefore, the FHC shines a helpful light on the situation, however, further discussion is needed by regulatory groups to determine what fairness measures should be weighed relatively more or less heavily for their context of interest. But knowing at minimum where companies lie along the spectrum through benchmarking procedures is a first step in enforcing quantitative fairness measures similar to performance metrics such as accuracy which are highly valued during the evaluation stage. Implications on regulatory actions and needs are discussed next.

4.3 Enabling Regulators and Going Public

With respect to governance and regulatory relevance, publishing a framework which can be used as an assessment template may increase regulators' attention to fairness in ML and, more importantly, the policies needed to ensure it. While standard metrics exist for other performance and functionality-based metrics (e.g., accuracy, profit and safety), similar quality assurance metrics for meeting ethical expectations are still limited. With the aid of FHC, regulators can shift the discussion towards implementing actionable measures for accountability with the use of a validated and transparently outlined methodology which companies (and auditors) can commit to. This is critical in a space suffering from vague

definitions and generic principles, all the while having grand expectations. This work only further emphasizes the need for urgent governance, which can differentiate two (sometimes competing) priorities of auditing for system effectiveness versus societal harm and optimize for an alternative goal centered around social and ethical values. Furthermore, this hints that an internal governance structure may soon be required for the evaluation of ML-based systems on the basis of ethical compliance. This alludes to new reality where these results could be made public to allow for adequate comparison and enforcement internally or by regulators to rectify any noncompliance, similar to the financial audits required by public companies. For example, ordering companies to annually (or more frequently) publish their fairness results and bias mitigation methodologies in a homogenous manner and take required action to reach certain industry-based thresholds could become a tangible possibility.

4.4 Unveiling of Protected Attributes

Finally, it is important to note that prior to having these expectations in place, ensuring accurate and representative outcomes would first require the unblinding of protected attributes. Collecting and having access to sensitive attributes such as protected class membership remains a serious challenge in practice, especially in traditionally highly regulated sectors such as credit and employment (Bogen et al., 2020; Holstein et al., 2018; Kallus et al., 2019). Access to sensitive attributes, beyond just proxies, are needed by industry practitioners in order to identify and estimate disparities in the first place, therefore making it fundamental to the cause for greater accountability (Dwork et al., 2011). The lack of access to this data is a real and grave concern as judgments drawn from proxy use is more vulnerable to criticism, can mislead decision making and ultimately restrict the reach and strength of future policy implications.

5 Closing Remarks

This paper makes several contributions. First, it articulates why and when current observational fairness criteria fall short of industry needs, thus bringing formalization to what was previously a jumble of notions scattered throughout literature. Second, this paper exposes the gap between the current state of fairness knowledge and the needs of industry practitioners and frames the problem in a way that indicates the lack of universal standard being a major pain point. Finally, this paper puts forward the Fairness Health Check, which can be used to manage fairer algorithm-assisted decision making. Given the complexity of this domain, this paper has shown that an effective fairness monitoring system based on a common set of indicators, which can be easily embedded in an established workflow, is key. These indicators can be used to identify strong or weak aspects of a model's fairness situation and facilitate inter-model or intertemporal comparisons while eliminating as much subjectivity as possible. Along

with monitoring progress, it plays a vital role in providing internal education on the topic, encouraging policy changes and nudging companies to take future action.

In closing, this work hopes to encourage practitioners to further build on this work and advance the nascent but promising domain of internal auditing discourse. It's important to remember that discrimination and bias did not appear out-of-the-blue with AI, but with accelerating technology cycles and infinitely scalable resources, the potential of AI to amplify and propagate these existing risks uncontrollably raises a fire alarm. The hope is to spark a critically relevant and timely discourse to ensure that theoretical considerations are properly translated into (future) norms in alignment with the prescribed AI principles and practical recommendations for building a more fair, inclusive society for all.

Author's Note

This whitepaper represents a work in progress with ideas based on my research until December 2021. Since then, I have pondered questions relating to how we can decisively select which fairness metrics are most relevant to the application at hand and who (and consequently how) we should shape standardized auditing practices in the future. I would like to emphasize that I do not believe simply measuring all types of fairness metrics and just "averaging across them" is the answer to this question as it would produce irresponsible results. Rather, this template is introduced as a first step in helping practitioners begin discussions around notions of fairness and bias and enabling organizations to identify the range, or maximum possible bounds, of disparity that exist in their data and algorithm as a starting point to developing a more holistic response. This is not meant to encourage a box-ticking activity, but rather to be used as an informational tool making up one component of a broader fairness strategy. By asking practitioners to assess the outcomes and implications of each type of fairness notion in relation to their specific ecosystem and particular consumer groups, we hope this will encourage responsible AI operationalization and ethical awareness for both experts and non-experts by highlighting the gaps where further attention is needed.

I welcome any feedback and would like to include a "call for action" for partners to form a working group as part of a discovery project in the next phase (industry players, institutional members, consumer groups, policy makers, regulators and others are welcome). This is an evolving conversation and I hope to continue adding to this work, specifically around setting industry-wide standards on how to define, measure and eventually regulate fairness in the Open Banking environment.

Contact: Savina Kim | PhD Candidate at the University of Edinburgh & Edinburgh Futures Institute | savina.kim@ed.ac.uk

Bibliography

Bakalar, C., Barreto, R., Bergman, S., Bogen, M., Chern, B., Corbett-Davies, S., Hall, M., Kloumann, I., Lam, M., Candela, J. Q., Raghavan, M., Simons, J., Tannen, J., Tong, E., Vredenburgh, K., & Facebook, J. Z. (2021). *Fairness On The Ground: Applying Algorithmic Fairness Approaches to Production Systems*. https://arxiv.org/abs/2103.06172v2

Bellamy, R. K. E., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., Zhang, Y., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., & Mehta, S. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4–5). https://doi.org/10.1147/JRD.2019.2942287

Berk, R. (2019). Accuracy and Fairness for Juvenile Justice Risk Assessments. *Journal of Empirical Legal Studies*, *16*(1), 175–194. https://doi.org/10.1111/JELS.12206

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods and Research*, *50*(1), 3–44. https://doi.org/10.1177/0049124118782533

Binns, R. (2019). On the Apparent Conflict Between Individual and Group Fairness. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 514–524. https://doi.org/10.1145/3351095.3372864

Bogen, M., Rieke, A., & Ahmed, S. (2020). Awareness in practice: Tensions in access to sensitive attribute data for antidiscrimination. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 492–500. https://doi.org/10.1145/3351095.3372877

Brennan, T., & Oliver, W. L. (2013). The Emergence of Machine Learning Techniques in Criminology: Implications of Complexity in our Data and in Research Questions Brennan and Oliver Forecasting Criminal Behavior. *Criminology and Public Policy*, *12*(3), 551–562. https://doi.org/10.1111/1745-9133.12055

Broek, E. van den, Sergeeva, A., & Huysman, M. (2019). Hiring Algorithms: An Ethnography of Fairness in Practice. *ICIS 2019 Proceedings*. https://aisel.aisnet.org/icis2019/future_of_work/future_work/6

Buolamwini, J. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification *. *Proceedings of Machine Learning Research*, *81*, 1–15.

Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Artificial Intelligence and Law*, *25*, 5–27. https://arxiv.org/abs/1610.07524v1

Corbett-Davies, S., Goel, S., Chohlas-Wood, A., Chouldechova, A., Feller, A., Huq, A., Hardt, M., Ho, D. E., Mitchell, S., Overgoor, J., Pierson, E., & Shroff, R. (2018). *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*. https://arxiv.org/abs/1808.00023v2

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *Part F129685*, 797–806. https://doi.org/10.1145/3097983.309809

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness Through Awareness. *ITCS 2012 - Innovations in Theoretical Computer Science Conference*, 214–226. https://doi.org/10.1145/2090236.2090255

E. O. of the President. (2016). *Big data: A report on algorithmic systems, opportunity, and civil rights*. Https://Obamawhitehouse.Archives.Gov/Sites/Default/Files/Microsites/Ostp/2016_0504_data_discrimination.Pdf.

Ensign, D., Frielder, S. A., Neville, S., Scheidegger, C., Venkatasubramanian, S., Mohri, M., & Sridharan, K. (2018). Decision making with limited feedback: Error bounds for predictive policing and recidivism prediction *. *Proceedings of Machine Learning Research*, *83*, 1–9.

Green, B., & Hu, L. (2018). *The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning*.

Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. *Undefined*, *2019-January*, 2122–2131. https://doi.org/10.24251/HICSS.2019.258

Grother, P., Ngan, M., & Hanaoka, K. (n.d.). *Face Recognition Vendor Test (FRVT), Part 3: Demographic Effects*. https://doi.org/10.6028/NIST.IR.8280

Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*, 3323–3331. https://arxiv.org/abs/1610.02413v1

Holstein, K., Vaughan, J. W., Daumé, H., Dudík, M., & Wallach, H. (2018). Improving fairness in machine learning systems: What do industry practitioners need? *Conference on Human Factors in Computing Systems - Proceedings*, 16. https://doi.org/10.1145/3290605.3300830

Hu, L., & Chen, Y. (2020). Fair classification and social welfare. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 535–545. https://doi.org/10.1145/3351095.3372857

Hutchinson, B., & Mitchell, M. (2018). 50 Years of Test (Un)fairness: Lessons for Machine Learning. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 49–58. https://doi.org/10.1145/3287560.3287600

Kallus, N., Mao, X., & Zhou, A. (2019). Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination. *Management Science*. https://doi.org/10.1287/mnsc.2020.3850

Kuhlman, C., Jackson, L., & Chunara, R. (2020). *No computation without representation: Avoiding data and algorithm biases through diversity*. https://arxiv.org/abs/2002.11836v1

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual Fairness. *Advances in Neural Information Processing Systems*, *30*. https://obamawhitehouse.archives.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data-

Lee, M. S. A., & Singh, J. (2020). The Landscape and Gaps in Open Source Fairness Toolkits. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.3695002

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). Delayed Impact of Fair Machine Learning. *IJCAI International Joint Conference on Artificial Intelligence*, *2019-August*, 6196–6200. https://doi.org/10.24963/ijcai.2019/862

Marjanovic, O., Cecez-Kecmanovic, D., & Vidgen, R. (2018). Algorithmic pollution: Understanding and responding to negative consequences of algorithmic decision-making. *IFIP Advances in Information and Communication Technology*, *543*, 31–47. https://doi.org/10.1007/978-3-030-04091-8_4

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6). https://doi.org/10.1145/3457607

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2021). Operationalising AI ethics: barriers, enablers and next steps. *AI and Society*, *1*, 1–13. https://doi.org/10.1007/S00146-021-01308-8/FIGURES/7

Narayanan, A. (2018). Tutorial: 21 fairness definition and their politics. *ACM FAT* (Fairness, Accountability and Transparency)*.

Neil Vigdor. (2019). *Apple Card Investigated After Gender Discrimination Complaints*. The New York Times. https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html

O'Neil, C., & Gunn, H. (2020). Near-term artificial intelligence and the ethical matrix. *Ethics of Artificial Intelligence*, 237–270. https://doi.org/10.1093/OSO/9780190905033.003.0009

ProPublica. (2016, May 23). *Machine Bias*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–481. https://doi.org/10.1145/3351095.3372828

Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. *AIES 2019 - Proceedings of the 2019*

AAAI/ACM Conference on AI, Ethics, and Society, 429–435. https://doi.org/10.1145/3306618.3314244

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020a). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020b). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing Algorithms : Research Methods for Detecting Discrimination on Internet Platforms. *Undefined*.

Shivayogi, P. (2013). Vulnerable population and methods for their safeguard. *Perspectives in Clinical Research*, *4*(1), 53. https://doi.org/10.4103/2229-3485.106389

Sokolovska, A., & Kocarev, L. (2018). Integrating Technical and Legal Concepts of Privacy. *IEEE Access*, *6*, 26543–26557. https://doi.org/10.1109/ACCESS.2018.2836184

Verma, S., & Rubin, J. (2018). Fairness Definitions Explained. *IEEE/ACM International Workshop on Software Fairness*, *18*. https://doi.org/10.1145/3194770.3194776

Whittlestone, J., Alexandrova, A., Nyrup, R., & Cave, S. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 195–200. https://doi.org/10.1145/3306618.3314289

Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation & Governance*, *12*(4), 505–523. https://doi.org/10.1111/REGO.12158

Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2016). Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. *26th International World Wide Web Conference, WWW 2017*, 1171–1180. https://doi.org/10.1145/3038912.3052660