Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics

Anonymised for review

March 23, 2021

Abstract

There is growing concern that decision-making informed by machine learning (ML) algorithms may unfairly discriminate based on personal demographic attributes, such as race and gender. Scholars have responded by introducing numerous mathematical definitions of fairness to test the algorithm, many of which are in conflict with one another. However, these reductionist representations of fairness often bear little resemblance to real-life fairness considerations, which in practice are highly contextual. Moreover, fairness metrics tend to be implemented in narrow and targeted toolkits that are difficult to integrate into an algorithm's broader ethical assessment. In this paper, we derive lessons from *ethical philosophy* and *welfare economics* as they relate to the contextual factors relevant for fairness. In particular we highlight the debate around acceptability of particular inequalities and the inextricable links between fairness, welfare and autonomy. We propose Key Ethics Indicators (KEIs) as a way towards providing a more holistic understanding of whether or not an algorithm is aligned to the decision-maker's ethical values.

Introduction

Algorithms are increasingly used to inform critical decisions in across high-impact domains, from credit risk evaluation to hiring to criminal justice. These algorithms are using more data from non-traditional sources and employing advanced techniques in machine learning (ML) and deep learning (DL) that are often difficult to interpret. The result is rising concern that these algorithmic predictions may be misaligned to the designer's intent, an

organisation's legal obligations, and societal expectations, such as discriminating based on personal demographic attributes. In response, there has been a proliferation of literature on algorithmic fairness aiming to quantify the deviation of their predictions from a formalised metric of equality between groups (e.g. male and female). Dozens of notions of fairness have been proposed, prompting efforts (Verma and Rubin, 2018) to disentangle their differences and rationale.

In line with this, a number of fairness *toolkits*[1] have been introduced to test the algorithm's predictions against various fairness definitions. The fairness toolkit landscape so far reflects the reductionist understanding of fairness as mathematical conditions, as the implementations rely on narrowly defined fairness metrics to provide "pass/fail" reports. These toolkits can sometimes give practitioners conflicting information about an algorithm's fairness, which is unsurprising given that it is mathematically impossible to meet some of the fairness conditions simultaneously (Kleinberg, Mullainathan, and Raghavan, 2016). This is reflective of the conflicting visions of fairness espoused by each mathematical definition and the underlying ethical assumptions (Binns, 2020).

A recent paper surveying the fairness toolkit landscape (Lee and Singh, 2020) found there were significant gaps between ML practitioner needs and the toolkits' features, especially regarding means that helped practitioners account for the contextual specifics of their use case – one practitioner commenting the toolkits "make everything look clear-cut, which it really isn't 'in the wild' " (Lee and Singh, 2020). Other studies involving ML practitioners have similarly emphasised the need for domain-specific and contextual factors to be closely considered to improve algorithmic fairness (Veale, Van Kleek, and Binns, 2018). In particular, in many domains, practitioners claim that fairness cannot be understood in terms of well-defined quantitative metrics (Holstein, Wortman Vaughan, Daumé III, Dudik, and Wallach, 2019).

This disconnect between real-world needs and axiomatic fairness definitions is not new. Hutchinson and Mitchell (2019) warn of the gap between the unambiguous formalisation of fairness metrics and the contextual and practical needs of society, politics, and law. They compared the recent surge in ML fairness research to literature from the 60s and 70s, which fizzled as "no statistic that could unambiguously indicate whether or not an item is fair was identified. There were no broad technical solutions to the issues involved in fairness" (Cole, 1973). From a legal standpoint, the approach in

[1]For example: IBM Fairness 360 (Bellamy, Dey, Hind, Hoffman, Houde, Kannan, Lohia, Martino, Mehta, Mojsilovic, et al., 2018), UChicago Aequitas tool (Saleiro, Kuester, Stevens, Anisfeld, Hinkson, London, and Ghani, 2018))

automating "fairness testing" appears incompatible with the requirements of EU non-discrimination law, which relies heavily on the context-sensitive, intuitive, and ambiguous evidence (Wachter, Mittelstadt, and Russell, 2020).

Fairness toolkits aim to be widely accessible, drawing attention to fairness considerations, and encouraging and supporting practitioners to consider, assess (and therefore mitigate) their algorithms in leading to unfair outcomes. However, without a consideration of the *relevant context* in the socio-technical system surrounding the algorithm, these tools risk engendering false confidence in flawed algorithms. Different considerations come into play for each use case. That is, organisations should not rely solely on one-dimensional algorithmic fairness metrics to account for its ethical concerns. These narrow applications of fairness could mislead organisational strategy, risk management, and policies.

Towards this, in this paper we draw from literature in *ethical philosophy* and *welfare economics* to pinpoint the relevant contextual information that should be considered in an understanding of a model's ethical impact. We argue that any future development of fairness toolkits should be framed within a broader view of ethical concerns to ensure their adoption promotes a contextually appropriate assessment of each algorithm.

To this end, we propose a new approach using *Key Ethics Indicators (KEIs)* to provide a more holistic understanding of whether or not an algorithm is aligned to the decision-maker's values. Though resembling some previous work on domain-specific trade-off analyses in fairness metrics vs. public safety (Corbett-Davies and Goel, 2018) and vs. financial inclusion (Lee and Floridi, 2020), our paper generalises the steps required for a holistic ethical assessment.

Our contribution is two-fold: 1) the identification of *relevant contextual factors* for fairness as drawn from ethical philosophy and welfare economics and 2) the proposal of a "Key Ethics Indicator" approach for a more comprehensive understanding of an algorithm's potential impact.

1 Definitions

We start by defining key terms: ethics, justice, fairness, equality, discrimination, and protected characteristics. This will frame our subsequent discussions on the contextual considerations for algorithmic ethics beyond what can be assessed using fairness metrics. While these dimensions the terms cover do not comprehensively cover all relevant aspects of algorithmic ethics, they clearly demonstrate the limitations of mathematical fairness formalisations in capturing necessary information about the algorithmic system.

As many organisations have launched initiatives to establish ethical principles, "AI ethics" definitions may vary; however, Floridi and Cowls identify the five common themes across these sets of principles: *beneficence, non-maleficence, autonomy, justice, and explicability* (Floridi and Cowls, 2019). We define algorithmic ethics along these five dimensions.

A study of proposed ethical principles finds that different countries' and organisations' understanding of *justice* varies for each document, from the elimination of discrimination to promoting diversity to shared prosperity (Floridi and Cowls, 2019). For the purpose of this paper, we distinguish between *justice* and *fairness* in accordance with legal and organisational science literature, with *justice* denoting adherence to the standards agreed upon in society (e.g. based on laws) and *fairness* as a related principle of an evaluative judgement of whether a decision is morally right (Goldman and Cropanzano, 2015).

In line with this definition, fairness is inherently subjective. The concept is based on the egalitarian foundation that humans are fundamentally *equal* and should be treated *equally*. However, how equality should be measured and to what extent it is desirable have been a source of debate in both philosophical ethics from a moral standpoint, and welfare economics from a market efficiency standpoint. What are the relevant criteria based on which limited resources should be distributed? For example, Aristotle wrote that if there are fewer flutes available than people who want to play them, they should be given to the best performers (Aristotle and Sinclair, 1962).

From a legal standpoint, *discrimination* refers to the notion that certain demographic characteristics, such as race and gender, should not result in a relative disadvantage of deprivation. Non-discrimination laws aim to not only prevent ongoing discrimination but also to change societal policies and practices to achieve more substantive equality – an aim which is described as incompatible with some fairness metrics (Wachter, Mittelstadt, and Russell, 2021). While legal analysis is outside the scope of this paper, we refer to protected characteristics as those commonly referenced and reflected in non-discrimination laws, such as race and ethnicity, gender, religion, age, disability, and sexual orientation, given these personal demographic features are central to discussions in algorithmic fairness literature. We also refer to 'direct' discrimination, which concerns differential treatment based on a protected characteristic and "indirect" discrimination, which represents an inadvertent negative impact on a protected group (Wachter et al., 2021).

2 Computer science literature

2.1 Fairness metrics

Existing mathematical definitions of fairness, while loosely derived from a notion of egalitarianism, should be calculated while keeping in mind the nuances and context-specificity present in philosophical discourse. We will walk through a use case: a lender building a model to predict a prospective borrower's risk of default on a loan. In this case, the False Positives (FP) represent lost opportunity (predicted default, but would have repaid), and the False Negatives (FN) represent lost revenue (predicted repayment, but defaulted).

The calculations of error rates used in the metrics are defined below, with some of the most commonly cited fairness definitions in Table 1:

- True Positive Rate (TPR) = TP/(TP + FN)
- True Negative Rate (TNR) = TN/(FP + TN)
- False Positive Rate (FPR) = FP/(FP + TN) = 1 – TNR
- False Negative Rate (FNR) = FN/(FN + TP) = 1 – TPR
- Positive Predictive Value (PPV) = TP/(TP+FP)

There are difficulties in deciding which metric is most appropriate for each use case (Lee and Floridi, 2020). Is a 3% increase in positive predictive parity preferable over a 5% increase in equal odds? Moreover that many of these metrics cannot be satisfied at the same time (Kleinberg et al., 2016), it is not intuitive on which metric best represents the lender's interests. These issues will be further discussed in §3, where we will link each fairness metric to its philosophical origin and address the gaps. In the next section, we challenge the types of inequalities that the fairness metrics assume are acceptable vs. unacceptable.

2.2 Acceptability of inequalities

First, we challenge the fairness metrics' assumed simplicity and separability of unacceptable bias by discussing the complexity of the debates on equality in ethical philosophy. Note that these metrics are aimed at a class of machine learning algorithms that are supervised, i.e. with a known outcome, and for classification purposes, i.e. for a discrete outcome (e.g. default vs. repayment) rather than a continuous outcome (e.g. amount repaid). These

Fairness metric	**Equalising**	**Intuition (Example)**
Maximise total accuracy	N/A	The most accurate model gives people the loan and interest rate they 'deserve' by minimising errors
Demographic parity, group fairness, disparate impact (Feldman, Friedler, Moeller, Scheidegger, and Venkatasubramanian, 2015)	Outcome	Black and white applicants have the same loan approval rates
Equal opportunity / false negative error rate balance (Hardt, Price, and Srebro, 2016)	FNR	Among applicants who are credit-worthy and would have repaid their loans, both black and white applicants should have similar rate of their loans being approved
False positive error rate balance / predictive equality (Chouldechova, 2017)	FPR	Among applicants who would default, both black and white applicants should have similar rate of their loans being denied
Equal odds (Hardt et al., 2016)	TPR, TNR	Meets both of above conditions
Positive predictive parity (Chouldechova, 2017)	PPV	Among credit-worthy applicants, the probability of predicting repayment is the same regardless of race
Positive class balance (Kleinberg et al., 2016)	Average probability of positive class	Both credit-worthy white and black applicants who repay their loans have an equal average probability score
Negative class balance (Kleinberg et al., 2016)	Average probability of negative class	Both white and black defaulters have an equal average probability score
Counterfactual fairness (Kusner, Loftus, Russell, and Silva, 2017)	Prediction in a counterfactual scenario in which the person had a different protected feature	For each individual, if he/she were a different race, the prediction would be the same
Individual fairness (Dwork, Hardt, Pitassi, Reingold, and Zemel, 2012)	Outcome for 'similar' individuals	For each individual, he/she has the same outcome as another 'similar' individual of a different race

Table 1: Fairness metrics

algorithms aim to identify the features that are associated with the outcome of interest. For example, one with higher income is more likely to be approved for a loan due to its association with higher ability to repay. In this case, differences in socioeconomic status is accepted as an inequality that is important to consider in the loan decision. Previously, scholars have made the distinction between "acceptable" vs. "unacceptable" inequalities based on legal precedents between "explainable" and "non-explainable" discrimination (Kamiran and Žliobaitė, 2013) based on Rawlsian philosophy between "relevant" and "irrelevant" features (Rawls, 1999). For example, income may be considered a "relevant" feature, and gender or race may be considered an "irrelevant" feature. The former should influence the algorithmic decisions, but the latter should not.

Scholars attempting to formalise these criteria into a mathematical definition of fairness have needed to address what type of equality is deemed to be fair. Some assume that any disparity in a given outcome metric is unacceptable (Gajane and Pechenizkiy, 2017): for example, loan approval rate for men and women should be the same. Others assume a level playing field (Gajane and Pechenizkiy, 2017): for example, there is no gender or racial discrimination in the real world that may affect the data. More recent work has taken a more nuanced stance, suggesting that the only features that should contribute to the outcome disparity are those that can be controlled by the individual, emphasising a distinction between the features driven by "effort" vs. "circumstances" (Heidari, Loi, Gummadi, and Krause, 2018). This is following the logic of Dworkin's theory of Resource Egalitarianism, no one should end up worse off due to bad luck, but rather, people should be given differentiated economic benefits as a result of their own choices (Dworkin, 1981).

Mathematical fairness formalisations must first determine which inequalities are "unacceptable." Some assume that all disparity in a given outcome metric is unacceptable, while others assume a level playing field (Gajane and Pechenizkiy, 2017), an assumption rarely met in societal challenges. Others emphasise the need to separate "effort" and "circumstances," suggesting that the only features that should contribute to the outcome disparity are those that can be controlled by the individual (Heidari et al., 2018). Another paper distinguishes between "benign" disparities and structural bias that should be corrected (Binns, 2020).

In reality, the layers of inequality between two individuals are intertwined, dynamic, and difficult to disentangle from one another. Consider the layers of inequality in Table 2. Two individuals may be unequal on several levels – in their level and type of talent, parents' socioeconomic status, behaviour, etc. – that may affect the target outcome of interest, whether it is credit-worthiness,

Types of inequality	Examples	Variable
Natural inequality	Disability at birth	Inequality 0
Socioeconomic inequality	Parents'/guardians' assets	Inequality 1
Talent inequality	Intelligence, skills, employment prospects	Inequality 2
Preference inequality	Saving behaviour, cultural prioritisation of values associated with economic opportunities	Inequality 3
Treatment inequality / societal discrimination (external)	Discrimination in job market and education system affecting income stability	Inequality 4

Table 2: Layers of inequality affecting the ground truth (*partial and indicative)

predicted performance at a job, or insurance risk. It is possible that the differences in the observed outcome are attributable to one or more of the above inequalities. Building an algorithm to predict the outcome could result in a faithful representation of these inequalities and the resulting replication and perpetuation of the same inequality through decisions informed by its predictions. However, which of the inequalities should be allowed to influence the model's prediction? We present the limitations of proposals thus far on how this question should be addressed.

2.2.1 Legally protected characteristics

The open source fairness toolkits often refer specifically to protected attributes in their assessment of fairness. For example, Fairness 360 defines protected attribute one that "partitions a population into groups whose outcomes should have parity. Examples include race, gender, caste, and religion" (Bellamy et al., 2018). While they acknowledge protected attributes may be application-specific, there is limited guidance on under what circumstances two groups should have parity in outcomes. In addition, how much disparity is acceptable in each use case and for each sub-group of interest? In order for the applications to be adopted across domain areas, it is important for practitioners to have a clear idea of what types of demographic features are acceptable vs. unacceptable to consider in an algorithm. However, often, computer science literature use these demographic features to assess fairness without challenging whether they are relevant to the decision at-hand.

Whether a disparity in fairness metrics between legally protected groups is fair depends on the context. Race and gender may be causally relevant in differential medical diagnosis (e.g. sickle cell anaemia, ovarian cancer) due to the different biological mechanisms in question. If the differences in outcome are causally related to the protected feature, the difference in decisions may be arguably fair. If a man has a higher income than a woman, he may receive a higher credit limit given his higher ability to repay.

2.2.2 Effort vs. circumstances

The suggestion to distinguish between the features driven by "effort" vs. "circumstances" in algorithmic fairness (Heidari et al., 2018) follows the logic of Dworkin's theory of Resource Egalitarianism, no one should end up worse off due to bad luck, but rather, people should be given differentiated economic benefits as a result of their own choices (Dworkin, 1981).

However, in reality, it is difficult to separate out what is within an individual's *genuine control*. For example, a credit market does not exist in a vacuum; while potential borrowers can improve their creditworthiness to a certain extent, e.g. by building employable skills and establishing a responsible payment history, it is difficult to isolate the features from discrimination in other markets, layers of inequality, and the impact of their personal history.

In addition to the challenge of defining what is within our control, some circumstances are necessary to take into account in a decision-making process. For example, one may not be in full control of one's income or education level, but they are crucial indicators of credit risk given they indicate greater job security. Socioeconomic and talent inequalities may be considered relevant in a credit risk evaluation algorithm, but they are not necessarily within our control.

2.2.3 Source of inequality

Scholars have also proposed that the *source* of inequality should determine which fairness metric is appropriate for each use case, i.e. whether the outcome disparity is explainable, justifiable, or benign or due to structural discrimination (Kamiran and Žliobaitė, 2013; Binns, 2020). Binns (2020) suggests group fairness metrics assumes disparities are benign, e.g. the loan approval difference between white and black applicants is solely due to their differences in ability to repay; statistical parity assumes structural bias that requires correction, e.g. historically, black applicants' risk have been inflated due to past discriminatory practices. However, in reality, there is rarely such a separation. For example, Lee and Floridi review the literature on U.S.

mortgage lending and suggest that there are many structural and statistical factors that lead the lenders to both over-estimate and under-estimate the risk of black borrowers (Lee and Floridi, 2020).

Any attempt to isolate the impact of discrimination from the impact of "benign" inequality needs to also consider the intersectional discrimination faced by those already marginalised in society (Crenshaw, 1989), e.g. the inter-connectivity of gender and racial discrimination (Collins, 2002). The boundary between what is an acceptable representation of existing inequalities and what is due to systematic discrimination and marginalisation of a group is challenging to ascertain.

Fleurbaey (2008) also cautions that "responsibility-sensitive egalitarianism" in welfare economics could be used to hastily justify inequalities and unfairly chastise the "undeserving poor" (Fleurbaey et al., 2008). He points out that the idea that people should bear the consequences of their choices is not as simple as it seems; it only makes sense when individuals are put in equal conditions of choice. Such an equality is not true in most systems. When one has fewer opportunities than another, one cannot be held fully responsible insofar as one's choice is more constrained.

2.2.4 Takeaways

The assumed clear and intuitive separation between acceptable and unacceptable inequalities, whether based on their source or the role of luck, rarely exists in real-life models. Not only is making the distinction impractical, the boundary itself is more controversially debated than it is often portrayed in algorithmic fairness literature, especially in computer science. The criteria for desirable equality depend on the philosophical perspective, which is ultimately a subjective judgement.

The decision on the target state—the way it ought to be—is an ethical decision with mathematically inevitable trade-offs between objectives of interest. Heidari et al. dismiss the distinction between relevant vs. irrelevant features in practice as out of scope for their paper: "Determining accountability features and effort-based utility is arguably outside the expertise of computer scientists" (Heidari et al., 2018). On the contrary, we argue that computer scientists and model developers must be actively engaged in the discussion on what layers of inequality should and should not be influencing the model's prediction, as this directly influences not only the model design and feature selection but also the selection of performance metrics.

3 Lessons from ethical philosophy on (in)equalities

Ethical philosophers have long debated whether equality is desirable and – if so – what type of equality people should pursue in society. Table 3 gives an example of philosophical perspectives and their perceptions of what types of inequality are acceptable. Formal equality of opportunity (EOP), or procedural fairness, posits that all opportunities should be equally open to all applicants (e.g. jobs, loans, etc.) based on a relevant definition of merit. However, in theory, this can be fully satisfied even if only a minority segment of a population (e.g. those with family wealth and connections) have realistic prospects of accessing the opportunity. In other words, as long as the opportunity is theoretically *available*, it is irrelevant whether it is *practically accessible*.

The Rawlsian fair EOP goes further to propose that any individuals with the same native talent and ambition should have the same prospects of success, requiring that all competitive advantage (e.g. parental efforts) be offset (Rawls, 1999). This is at odds with Lockean and libertarian ideals that assert the value of each person's freedom insofar as there is no harm to another (Nozick, 1974), which naturally extends to the right to ownership and capital. Rawls also proposes the Difference Principle as an exception: economic and social inequalities can only be justified if they benefit the most disadvantaged members of society (Rawls, 1999). These EOP principles are in contrast to the strict equality of outcome, condition, or welfare, which requires an equal distribution regardless of any relevant criteria.

Luck egalitarians hold that unchosen inequalities must be eliminated (Dworkin, 1981). Sen and Fleurbaey object on the grounds that luck egalitarians have no principled objection to a society in which, on a background of equal opportunities, some end up in poverty or as the slaves of others (Fleurbaey et al., 2008). They argue for a more substantive equality of "autonomy" that includes the full range of individual freedom.

Some have argued that what is important is not relative condition compared to other people, but rather, whether people have enough to have satisfactory life prospects (Walzer, 1983). Others have shifted the focus on the incremental gain of well-being of those who are worst-off (Parfit, 1991). Yet others have debated the foundations of desert, or what one deserves corresponding to his or her virtue (Kagan, 2014).

3.1 Ethical subjectivity of algorithmic fairness

As such, what types of inequality in outcome are fair is a philosophical and subjective debate with nuances and complexities insufficiently addressed in

Philosophical perspective	Acceptable inequalities	Unacceptable inequalities
Formal equality of opportunity / procedural fairness (Greenberg, 1987)	Any inequality as long as the opportunity was open to all	Treatment inequality
"Fair equality of opportunity" (Rawls, 1999, 2001)	Natural, talent, and preference inequalities	Socioeconomic, treatment inequalities
Rawlsian EOP + Difference principle (Rawls, 1999)	Natural, talent, and preference inequalities, plus any inequality benefiting the most disadvantaged society members in long-term impact	Socioeconomic, treatment inequalities
Equality of outcome / condition / welfare (Greenberg, 1987)	None - all members should get the exact same outcome	All
Luck egalitarianism (Dworkin, 1981)	Effort-based inequalities (e.g. preference)	Circumstances (e.g. natural inequality)
Equality of freedom / autonomy (Sen, 1992)	Inequality resulting in "genuinely free" choices	Any inequality hindering freedom
Sufficiency / Equality of capability (Walzer, 1983)	Any inequality as long as everyone is above the level of sufficiency	Any resulting in people falling below sufficiency levels
Prioritarianism (Scheffler, 1994; Parfit, 1991)	Any inequality reduction should prioritise resource allocation to those who are worst off	None as long as the worst off are prioritised
Desert (Kagan, 1999, 2014)	Any inequality based on what he/she "deserves"	Any inequality that does not equate to the person's deservingness

Table 3: Key philosophical perspectives on inequality

existing algorithmic fairness literature. What happens when faithfully representing the world as it is perpetuates an unfair state of affairs? This complicates the objective of machine learning, which is only reliable insofar as it is trained on data sets that reflects reality. To forecast sales, an algorithm learns from data representative of the company's customers. For example, online searches for "CEO" yield mostly images of white men, and online job postings may show high-income positions to men more frequently than women (Van Dam, 2019). This may result in a biased outcome, with men securing disproportionately high-paying jobs. However, this is reflective of the existing gender pay gap: in 2019, only 6.6% of Fortune 500 top executives were female, the highest proportion in history (Zillman, 2019). Continuing to under-represent women in search results may perpetuate the bias that CEOs are typically men. In this instance, some call for the "correction" of the bias to reflect judgements about the way the world *should* be, which is by nature an ethically influenced choice.

As previously stated, on the contrary to past scholars' arguments (Heidari et al., 2018), our position is that computer scientists and model developers cannot completely delegate this consideration to a third party, whether it is the regulator, business leader, or the risk function. Model developers must be engaged in the discussion on what layers of inequality should and should not be influencing the model's prediction in order to inform their decisions on model design, feature selection, and performance metric selection.

Overall, in formalising fairness, the decision-maker should be explicit on 1) which inequalities and biases exist that affect the outcome of interest and 2) on which of them should be retained and which of them should be actively corrected. This will be further addressed in our proposal of Key Ethics Indicators (KEIs). We next link some of the fairness metrics to the ethical philosophy that inspired them, pointing out the contextual considerations in the ethical philosophy that should be kept in mind alongside the fairness formalisations.

3.2 Linking ethical philosophy to algorithmic fairness

Existing mathematical definitions of fairness, while loosely derived from a notion of egalitarianism, should be calculated while keeping in mind the nuances and context-specificity present in philosophical discourse. Revisiting the fairness metrics from Table 1, this section will show link each metric to the ethical philosophy that inspired it, as well as addressing the gaps between the philosophical work and what is represented in the mathematical formula.

In discussing the entries of Table 4 in order: accuracy maximisation is prone to biases introduced in the model development lifecycle that may skew

Fairness metric	**Equalising**	**Philosophy**
Maximise total accuracy	N/A	Desert (Kagan, 1999, 2014)
Demographic parity, group fairness, disparate impact (Feldman et al., 2015)	Outcome	Strict egalitarianism (Equality of outcome / condition / welfare) (Greenberg, 1987)
Equal opportunity / false negative error rate balance (Hardt et al., 2016)	FNR	
False positive error rate balance / predictive equality (Chouldechova, 2017)	FPR	
Equal odds (Hardt et al., 2016)	TPR, TNR	
Positive predictive parity (Chouldechova, 2017)	PPV	"Fair equality of opportunity" (Rawls, 1999, 2001)
Positive class balance (Kleinberg et al., 2016)	Average probability of positive class	
Negative class balance (Kleinberg et al., 2016)	Average probability of negative class	
Counterfactual fairness (Kusner et al., 2017)	Prediction in a counter-factual scenario in which the person had a different attribute	David Lewis, cause and effect (Lewis, 1973)
Individual fairness (Dwork et al., 2012)	Outcome for "similar" individuals	Responsibility-sensitive egalitarianism (Fleurbaey et al., 2008)

Table 4: Fairness metrics and their philosophical origins

the predictions, which is especially problematic if the biases reflect patterns of societal discrimination, leading to "undeserved" outcomes contrary to the philosophy of desert. Demographic parity is problematic if there are legitimate rationale behind the unequal outcome (e.g. unequal income).

The equal opportunity metric, while it sounds attractively similar to Rawlsian EOP, fails to address discrimination that may already be embedded in the data (Gajane and Pechenizkiy, 2017). Discrimination may be crystallised in the data set due to biased data collection (e.g. selective marketing), biased data labelling (e.g. humans scoring male candidates as more competent), or biased human decisions feeding the system (e.g. if courts are more likely to find black defendants as guilty). Rawlsian EOP also assumes that inequalities in native talent and ambition may result in unequal outcomes, which is not addressed in the equalisation of false negative rates. Each group fairness metric, including equal odds, positive predictive parity, and positive / negative class balance, requires different assumptions about the gap between the observed space (features) vs. the construct space (unobservable variables): "if there is structural bias in the decision pipeline, no [group fairness] mechanism can guarantee fairness" (Friedler, Scheidegger, and Venkatasubramanian, 2016). This is supported in a critique of existing classification parity metrics, in which the authors conclude that "to the extent that error metrics differ across groups, that tells us more about the shapes of the risk distributions than about the quality of decisions" (Corbett-Davies and Goel, 2018). In many domains in which there are concerns over unfair algorithmic bias, including credit risk and employment, there has often been a documented history of structural and societal discrimination, which may affect the underlying data through biases previously discussed.

The challenge of individual fairness approach is: how to define "similarity" that is, for example, independent of race (Kim, Reingold, and Rothblum, 2018). When the predictive features are also influenced by protected features, designation of a measurement of "similarity" cannot be independent of those protected features. For example, what proportion of gender income disparity is due to structural employment discrimination as opposed to job preferences? Some scholars have attempted to incorporate active corrections for racial inequality into metrics of similarity (Dwork et al., 2012), but this depends heavily on the assumption that the inequality due to racial discrimination can be isolated from other sources of inequality.

While counter-factual fairness metric provides an elegant abstraction of the algorithm, the causal mechanisms, e.g. of a default on a loan or on insurance risk, are typically not well understood. It is also difficult to isolate the impact of one's protected feature, e.g. race, on the outcome, e.g. risk of default, from the remaining features. The approach is also sensitive to

unmeasured confounding variables, which may add additional discriminatory bias (Kilbertus, Ball, Kusner, Weller, and Silva, 2019).

In all, these metrics do not give any information on which layers of inequalities they are attempting to correct, which risks over- or under-correction. A deeper engagement with the ethical assumptions being made in each model is necessary to understand the drivers of the unequal outcomes. What types of inequalities are acceptable depends on the context of the model. Our KEI approach will account for such context-specificity of what inequalities are acceptable.

4 Lessons from welfare economics

Referring back to our definition of algorithmic ethics, justice is only one of five dimensions (*beneficence, non-maleficence, autonomy, justice, and explicability*), with fairness as a key principle related to justice. We derive lessons from literature on welfare economics to demonstrate the interconnectedness of fairness and welfare (*beneficience* and *non-maleficence*) and liberty (*autonomy* and *explicability*). By focusing narrowly on the fairness metrics, which quantify the redistribution of the target outcome, a decision-maker may overlook the key considerations of the impact on the stakeholders' welfare and autonomy. Because of the challenge in quantifying the relevant biases and disentangling them from the outcome of interest, correcting for a bias carries the risk of increasing the inaccuracies of the predictions. Beyond the egalitarian perspective on the relative distribution of resources between individuals and groups, it is important to consider the aggregate impact of an algorithm on the society.

4.1 Welfare in algorithmic ethics: beneficence and non-maleficence

We use an example concerning credit risk evaluation to argue that fairness should be considered alongside welfare. In attempting to improve a fairness metric, a decision-maker may inadvertently forego an algorithm that leaves everyone better-off (beneficence) or may inadvertently harm the sub-group they are attempting to help to "level the playing field." Fairness metrics should not be taken at face value without an understanding of how they may affect other ethical objectives. Fairness toolkits that assess fairness in isolation risks misleading the decision-makers by giving the them incomplete information about whether their algorithm meets their ethical objectives.

From a welfare economic standpoint, a notion of fairness includes a consideration of well-being: from both utilitarian and libertarian perspectives, a fair reward principle maximises the sum total of individual well-being levels while legitimising redistribution that enhances the total outcome of individuals (Fleurbaey et al., 2008). This is not necessarily contradictory to the egalitarian perspectives discussed in ethical philosophy. In accordance with the Difference Principle, Rawlsian EOP Max-Min social welfare function should also maximise the welfare of those who are worst-off (Rawls, 1999). A model that results in financial harm of already-disadvantaged populations fails to meet the Rawlsian EOP criteria, even if the False Negative Rates are equalised as per the mathematical definition. Without consideration of the long-term impact on welfare, the fairness metrics fail to capture the full extent of ethical dilemma embedded in a model selection process.

Accuracy is often considered in trade-off with fairness (Kleinberg et al., 2016), but from an ethical standpoint, that accuracy may represent a key principle in beneficence or non-maleficence. For an example of beneficence, a "good" credit risk algorithm would lower the aggregate portfolio risk for the lender, enabling more loans to more people and giving them access to credit that is crucial to upward socioeconomic mobility. For an example of non-maleficence, the false positive rates (i.e. loans that were approved but defaulted) also contains information about whether unaffordable loans are granted. A lender should aim to minimise the borrower's financial difficulty, given the adverse effects of unaffordable debt on both the market level (causing instability and a "bubble") and for the borrower (Aggarwal, 2018).

The ethical principle of non-maleficence may be in direct conflict with fairness in some circumstances. Adding fairness constraints may end up harming the groups they intended to protect in the long-term (Liu, Dean, Rolf, Simchowitz, and Hardt, 2018). In the presence of a feedback loop, we need to consider not only providing a resource (a loan) to an applicant in a disadvantaged group, but also what happens as a result of that resource. If the borrower defaults, his/her credit score will decline, potentially precluding the borrower from future loans. It is important to view fairness, not in isolation at a moment in time, but rather, in the context of long-term objectives in promoting the customer's financial well-being. This is a part of the context we formalise in our Key Ethics Indicator proposal.

4.2 Liberty in algorithmic ethics: autonomy and explicability

Fairness should also be assessed within the context of how the algorithm affects human *liberty*, a subject in welfare economics that is relevant to the AI ethics principles of *autonomy* and *explicability*. Fleurbaey argues responsibility-sensitive egalitarianism in welfare economics should move away from "responsibility," which may overlook certain people's lack of freedom to choose alternatives, and towards "autonomy" (Fleurbaey et al., 2008). In other words, for there to be "true" equality, three conditions must be met: 1) a minimum level of autonomy is attained, 2) with a minimum level of variety and quality of options offered, 3) with a minimum decision-making competence (Fleurbaey et al., 2008). A comprehensive egalitarian theory of justice is not just about equalising available opportunities but also about providing adequate opportunities and making them *accessible.* As per our definition of AI ethics Floridi and Cowls, we define autonomy as the power to decide, striking a balance between the decision-making power humans retain and that which we delegate to artificial agents. We also define *explicability* as the combination of *intelligibility* (how it works) and *accountability* (who is responsible for the way it works). It complements the other four principles by helping us understand the good or harm an algorithmic system is actually doing to society, in which ways, and why Floridi and Cowls.

4.2.1 Autonomy: Liberty

In enforcing some of the stricter fairness conditions, decision-makers should be careful as to the potential impact this has on human autonomy. Luck egalitarians, for instance, have no principled objection to a society in which, on a background of equal opportunities, some end up in poverty or as the slaves of others (Fleurbaey et al., 2008) – this could violate fundamental human rights to freedom and result in undesirable levels of extreme societal inequality. Intervention is necessary when basic autonomy is at stake, and this should be a constraint on definition of fairness. Fleurbaey argues this is consistent with egalitarian welfare economics, as egalitarians should be concerned—not only with equality of opportunities—but also with the content of the opportunities themselves, with freedom as the leading principle in defining responsibility in social justice (Fleurbaey et al., 2008).

By focusing on equality of opportunities, one may dismiss the differences in preferences as driven by choice and thus irrelevant. However, Fleurbaey argues that the ex post inequalities due to differences in preferences are also a target for intervention on the grounds of improving the range of choices

to suit everyone's preferences. If more women prefer lower-paid positions than men, what is problematic is not only the societal and environnmental conditioning that questions whether this is a genuine preference, but also the unfair advantage that attaches to these jobs – a differential value of the "menu" of options for women than for men because of their preferences (Fleurbaey et al., 2008). Considerations of fairness and the associated policy response must operate at the level of the menu, rather than distribution of jobs themselves.

4.2.2 Autonomy: Forgiveness

Fleurbaey also discusses a concept that is not addressed in algorithmic fairness literature: forgiveness. He argues that the ideal of freedom and autonomy contains the idea of "fresh starts": in absence of cost to others, it is desirable to give people more freedom and a greater array of choices in the future (Fleurbaey et al., 2008). This is in conflict with the "unforgiving conception of equality of opportunities" that ties individuals to the consequences of one's choices (Fleurbaey et al., 2008). In many countries, lenders are restricted in their access to information about borrowers' past defaults; for example, many delinquencies are removed from U.S. credit reports after seven years (Elul and Gottardi, 2015). Forcing a lender to ignore information about past behaviour may reduce the accuracy of its default prediction model, and it may be "unfair" by some definitions by putting those who have made more responsible financial decisions on equal level as those who have not; however, it is widely accepted practice to ensure that one decision does not have a disproportionate impact of limiting one's access to credit for good. A more complete coverage of fairness and justice, therefore, should go beyond redistribution of outcome features and consider the impact on individual welfare, autonomy, and freedom.

4.2.3 Autonomy: Vulnerability

Autonomy in rational decision-making also falters as an ethical objective when there is a significant asymmetry of power and information between two parties. Contractarian perspectives on fairness assumes two equal entities exchanging one resource for another(Gauthier, 1986).

Those with limited autonomy include vulnerable customers. When an algorithm targets and manipulates those with no other options, they do not have the autonomy to enter into the contract, whether or not the contract is fair. Payday loans and check cashing industry in the US targets those who cannot access traditional financial services, often due to their illegal

immigration status or long working hours that do not provide a break while a bank is open for business, entrapping the most vulnerable groups into an unbreakable cycle of debt with unaffordable interest rates (Prager et al., 2009). While the interest rate may not necessarily be unfair (it may in some cases be proportional to the likelihood of an individual's repayment), it is ethically undesirable. The same principle applies to marketing insurance products to those with recent bereavement or the sale of complex financial instruments to someone without the capability of understanding their risks.

Another group is those with "thin" files, with a lack of or sparse credit history. There has been a movement to use "alternative data" or non-traditional data sources that do not directly relate to the borrower's ability to repay. One of the most extreme cases is the use of Internet browsing history, location, and payment data to calculate credit risk (Koren, 2016). The justification is often that this increases financial inclusion for those without alternate means to access credit. However, this requires the lender access to more data from the currently unbanked populations, disproportionately forcing them to give up more of their privacy than those with existing credit histories. It also provides additional risk of discrimination, as the non-traditional data sources are likely to be closely intertwined with personal characteristics. Location and social media data are more likely to reveal an individual's race and gender than credit history. While Kenya's poor were among the first to benefit from digital lending applications, they have led to a predatory cycle of debt the borrowers describe as a new form of slavery, between the endless nudges to borrow, the lenders' control over a vast archive of user data, and the ballooning interest payments (Donovan and Park, 2019). This double-standard of privacy between the unbanked and banked violates the equal rights of individuals to privacy and self-determination. While there may be an exchange of access to credit and personal data (e.g. if an individual gives consent to a personality test or access to his/her social media profile), there should be a protection of their right to privacy.

Fairness overall must be considered in the context of the impact on individual human rights – going beyond the equality of available opportunities, empowering human freedom and autonomy to ensure *accessibility* of these opportunities. Computer scientists can learn from the welfare economists' consideration of autonomy as a crucial component of egalitarian perspectives on fairness.

4.2.4 Explicability

Welfare economics is built on the assumption of rational, free agents, which is shared in Kantian ethical philosophy (Kant and Gregor, 1996). This has

been applied to medical ethics to mandate that a patient be able to make a fully informed decision on whether or not to receive treatment (Eaton, 2004). Similarly, in algorithmic decision-making, individuals consenting to the usage of their data should fully understand how the data will be used. When humans employ autonomous systems, they cede, at least provisionally, some of their own autonomy (decision-making power) to machines (Floridi and Cowls, 2019). Respecting human autonomy thus becomes a matter of ensuring that both the decision-making authority and the subject of the decision retain enough autonomy to safeguard their well-being.

In order to incorporate the algorithm into rational decision-making, it is important to understand how the algorithm reached its prediction or recommendation. Due to the relatively limited interpretability of ML, "explainable AI" (xAI) is an ongoing area of research (Xu, Uszkoreit, Du, Fan, Zhao, and Zhu, 2019). There is often a trade-off between accuracy of an algorithm and its explainability, as complex phenomena are better represented by complex, "black-box" models than simple and interpretable models. This may, in turn, represent a trade-off between explainability (and thus a decision-maker's capability for reasoning) and any beneficence afforded by the increase in accuracy and model performance. In some use cases, e.g. film recommendations, accuracy may outweigh the need for explanations. The explanations may vary based on the target of the explanation, e.g. customer, regulator, domain experts, or system developers (Arya, Bellamy, Chen, Dhurandhar, Hind, Hoffman, Houde, Liao, Luss, Mojsilović, et al., 2019). It is important to understand the interplay between an algorithm's explanation and its perceived fairness. There may be a number of possible explanations for any given decision, and the techniques for xAI alone do not detect or correct unfair outcomes. However, the explanations may help identify potential variables that are driving the unfair outcomes, e.g. if pricing varies for female-dominated professions compared to male-dominated professions, the model may be relying on occupation for its prediction, which acts as a proxy for gender.

While fairness formalisations may provide a simple methodology for model developers to incorporate metrics relevant to equalisation of outcomes between groups and individuals, they do not provide a holistic view of the important debates on what fairness means, as they are discussed in ethical philosophy and welfare economics. The narrow definition of unfair bias in each of these metrics only provides a partial snapshot of what inequalities and biases are affecting the model and does not consider the long-term and big-picture ethical goals beyond this equalisation.

5 Proposed method: Key Ethics Indicators

In this final section, we propose a new approach that moves away from attempts to define fairness mathematically, and instead, gain a more holistic view of the ethical considerations of a model. Due to the subjectivity of fairness metrics, it may be challenging to select one over another. Rather than these general metrics, decision-makers should create a customised measurement of what "fair" looks like in each model. In addition, fairness should not be considered in isolation from the related ethical goals. The interaction between fairness and other values - e.g. welfare, autonomy, and explicability - should be taken into account in this analysis.

Contrary to claims otherwise (Heidari et al., 2018), the roles and responsibilities of a developer are necessarily intertwined with the role of the expert or business stakeholder, as the ethical and practical valuations of what "success" looks like in the model directly influences the algorithm design, build, and testing. It is important to have active engagement from the beginning between the developer and the subject matter expert to try to understand which inequalities should influence the outcome and how to address the inequalities that should not play a role in the prediction. This process requires engagement from all relevant parties, including the business owner and the technical owner, with potential input from regulators, customers, and legal experts.

Relying solely on the out-of-the-box fairness definitions as implemented in fairness toolkits would fail to capture the nuanced ethical trade-offs. For a decision-maker, it is important to devise customised success metrics specific to the context of each model, which as we described, involves considering welfare (beneficence, non-maleficence), autonomy, fairness, and explicability. This can be done in a following process:

1. Define "success" from an ethical perspective. What is the benefit of a more accurate algorithm to the consumer, to society, and to the system? What are the potential harms of false positives and false negatives? Are there any fundamental rights at stake?

2. Identify the layers of inequality that are affecting the differences in outcome

3. Identify the layers of bias

4. Devise an appropriate mitigation strategy. This may require changes to data collection mechanism or to existing processes, rather than a technical solution.

5. Operationalise these objectives into quantifiable metrics, build multiple models and calculate the trade-offs between the objectives covering all ethical and practical dimensions.

6. Select the model that best reflects the decision-maker's values and relative prioritisation of objectives.

We now elaborate each of these steps, in turn.

5.1 Define success

For each use case, there are unique considerations on what is considered a "successful" model, which are unlikely to be captured in a single mathematical formula. In credit risk evaluation, for example, three key objectives from ethical, regulatory, and practical standpoints are: 1) allocative efficiency: a more accurate assessment of loan affordability protects both the lender and the customer from expensive and harmful default, 2) distributional fairness: increasing access to credit to disadvantaged borrowers, including "thin-file" borrowers and minority groups, 3) autonomy: both increased scope of harm due to identity theft and security risk and due to the effects of ubiquitous data collection on privacy (Aggarwal, 2020). Here, a successful credit risk model would achieve all three objectives. By contrast, in algorithmic hiring, success metrics may include employee performance, increased overall diversity among employees and in leadership, and employee satisfaction with the role. It is important to identify all the objectives of interest, such that any trade-offs between them may be easily identified, allowing for a more holistic view of algorithmic ethics.

5.2 Identify sources of inequality

As previously discussed, due to the complex and entangled sources of inequalities and bias affecting an algorithm, there is no simple mathematical solution to unfairness. It is important to understand what types of inequality are acceptable vs. unacceptable in each use case. Table 2 presents different layers of inequality. Considering a credit risk evaluation, socioeconomic and talent inequalities may be considered relevant: if a man has a higher income than a woman, he may receive a higher credit limit given his higher ability to repay; higher education level and expertise in a high-demand field may indicate greater job security. Forcing the decision-maker to look beyond the legally protected characteristics to identify the inequalities that are acceptable and relevant and those that are not helps better identify the sub-groups that are at risk of discrimination.

5.3 Identify sources of bias

In addition to the inequalities discussed above, there may be biases in the model development lifecycle that exacerbate the existing inequalities between two groups. The challenge is that in many cases, the patterns associated with the target outcome are also associated with one's identity, including race and gender.

Suresh and Guttang (2020) have recently grouped these types of biases into 6 categories: historical, representation, measurement, aggregation, evaluation, and deployment. Historical bias refers to past discrimination and inequalities, and the remaining five biases, displayed in Table 5, align to the phases of the model development lifecycle (data collection, feature selection, model build, model evaluation, and productionisation) that may inaccurately skew the predictions. By understanding the type of bias that exists, the developer can identify the phase in which the bias was introduced, allowing him or her to design a targeted mitigation strategy for each bias type.

Table 5 gives examples in racial discrimination in lending processes to demonstrate each type of bias. For a practical tool in identifying unintended biases in these six categories, see: Lee and Singh (2021). Crucially, they point out that effective bias mitigation addresses the bias at its source, which may involve a non-technical solution. For example, bias introduced through the data collection process may require a change in marketing strategy.

5.4 Design mitigation strategies

The mitigation strategy depends on whether we believe the inequalities in Table 2 and the biases in Table 5 need to be actively corrected to rebalance the inequalities and bias. It is important to understand the source of the bias in order to address it.

There have been existing methods proposed for *pre-processing*, removing bias from the data before the algorithm build, *in-processing*, building an algorithm with bias-related constraints, and *post-processing*, adjusting the output predictions of an algorithm. However, these methods presume that inequalities in Table 2 and the biases in Table 5 are known and can be quantified and surgically removed. How do we isolate the impact of talent and preference inequalities on income from the impact of discrimination? The attempt to "repair" the proxies to remove the racial bias has been shown to be impractical and ineffective when the predictors are correlated to the protected characteristic; even strong covariates are often legitimate factors for decisions (Corbett-Davies and Goel, 2018).

Often, the solution to these biases is not technical because their sources

Types of bias	Examples	Variable
Representation bias	Limited marketing and outreach in high-minority neighborhoods	Bias 0
Measurement bias	Unequal treatment in the lending process associated with race leads to mis-measurement of risk factors	Bias 1
Aggregation bias	There may be a difference in default frequency distribution between racial groups, which is poorly represented by a single model	Bias 2
Evaluation bias	The accuracy and precision metrics in default prediction vary across racial groups (e.g. lower confidence in predictions for minority borrowers)	Bias 3
Deployment bias	True outcome only known for accepted loans and unknown for denied loans	Bias 4

Table 5: Layers of bias resulting in inaccurate predictions (*partial and indicative)

are not inherent in the technique. Instead of looking for a mathematical solution, there may be productive ways of counteracting these biases with changes to the process and strategy. Examples are shown in Table 6.

Types of bias	**Variable**	**Example action**
Treatment inequality / societal discrimination (external)	Inequality 4	Identify a new feature to estimate income volatility associated with race
Representation bias	Bias 0	Change in marketing and outreach strategy to include more high-minority neighborhoods
Measurement bias	Bias 1	Employee training on subconscious bias, standardized practice on which loan types are recommended based on prespecified relevant criteria
Deployment bias	Bias 2	Continuous monitoring and analysis of whether the decision boundary between rejection and acceptance is appropriate

Table 6: Possible actions to counteract biases (*partial and indicative)

While the mitigation strategies are important, they are unlikely to provide a complete solution to the problem of algorithmic bias and fairness. That is because—unlike the assumptions underlying fairness formalisations—it is often not feasible to mathematically measure and surgically remove unfair bias from a model, which is affected by inequalities and biases that are deeply entrenched in society and in the data.

Legal scholars have argued that traditional approach of scrutinising the inputs to a model is no longer effective due to the rising model complexity. Using Fair Lending law as an example, Gillis demonstrates that identifying which features are relevant vs. irrelevant fails to address discrimination concerns because combinations of seemingly relevant inputs may drive disparate outcomes between racial group (Gillis, 2020). Rather than focusing on identifying and justifying inputs and policies that drive disparities, Gillis argues, it is important to shift to an *outcome-focused* analysis of whether a model leads to impermissible outcomes (Gillis, 2020). Similarly, Lee and Floridi have proposed an approach to assess whether the outcome of a model is desirable (Lee and Floridi, 2020). For a more comprehensive analysis of whether a model meets the stakeholders' ethical criteria, it is important to look beyond the inputs and the designer's intent and assess the long-term and holistic outcome.

5.5 Operationalise Key Ethics Indicators (KEIs), calculate trade-offs between KEIs

Once "success" for a model has been defined at a high-level, the next step is to operationalise the ethical principles such that they are measurable. Similarly to how a company may define a set of quantifiable values to gauge its achievements using Key Performance Indicators (KPIs), there should be outcome-based, quantifiable statements from an ethical standpoint: Key Ethics Indicators (KEI), enabling developers to manage and track to what extent each model is meeting the stated objectives.

For example, Lee and Floridi estimate the impact of each default risk prediction algorithm on financial inclusion and on loan access for black borrowers (Lee and Floridi, 2020). They operationalise financial inclusion as the total expected value of loans under each model and minority loan access as the loan denial rate of black applicants under each model. In Figure 1 replicated from their work, they calculate the trade-offs between the two objectives for five algorithms, providing actionable insights for all stakeholders on the relative success of each model.

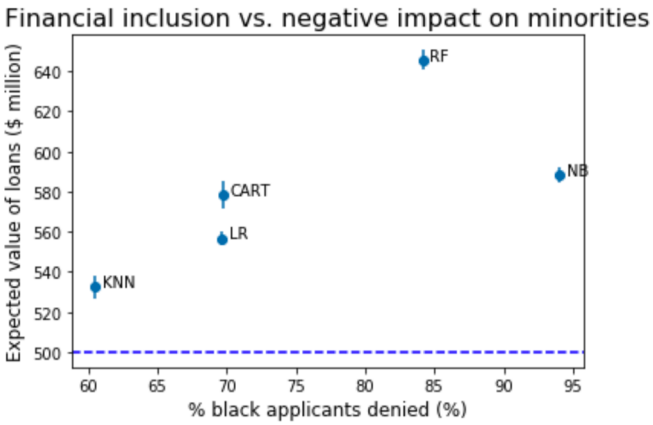

Figure 1: Replicated from Lee and Floridi (2020): Trade-off analysis

Context-specific KEIs can be developed for each use case. For example, in algorithmic hiring, employee satisfaction with a role may be estimated by attrition rates and employee tenure, employee performance may be measured through their annual review process, and diversity may be calculated across gender, university, region, age group, and race, depending on each organisation's objectives and values. Making explicit the ethical objectives in each use case would help decision-makers justify the use of any algorithm, which could in turn lead to the establishment of industry standards, informing best

practices, policy design, and regulatory activity.

5.6 Select a model and provide justifications

The trade-off analysis makes the ethical considerations clear. For example, in Figure 1, Lee and Floridi conclude that Random forest is better in absolute terms (in both financial inclusion and impact on minorities) than Naïve Bayes, but the decision is more ambiguous between CART and LR: while CART is more accurate and results in greater financial inclusion (equivalent of $15.6 million of loans, or 103 median-value loans), CART results in a 3.8 percentage points increase in denial rates for black loan applicants compared to LR. This quantifies the concrete stakes to the decision-maker who may decide on the model that is most suited to his or her priorities, customised to each use case.

One of the key benefits of the outcome-driven KEI trade-off analysis is that it provides interpretable and actionable insights into the decision-maker's values, which is especially important for complex machine learning algorithms in which the exact mechanism may not be transparent or interpretable. This could also provide valuable justification to the regulator on why a certain model was seen as preferable to all other reasonable alternatives. This may also help reduce the hesitation among decision-makers around the use of machine learning models due to their non-transparent risks, if the analysis shows they are superior to traditional rules-based models in meeting each of the KEIs. Suitable records of the decisions must be kept, ensuring the model and its design are *reviewable* (Cobbe, Lee, and Singh, 2021).

Conclusion

Implementations of fairness toolkits have predominantly implemented mathematical fairness definitions without locating their implications in overall algorithmic ethics. One of our contributions is to derive lessons from *ethical philosophy* and from *welfare economics* on what are the *contextual considerations* that are important in assessing an algorithm's ethics beyond what can be captured in a mathematical formula. For example, we refer to the debate in ethical philosophy on what constitutes acceptable vs. unacceptable inequalities. We also relate to the explicit consideration in welfare economics of welfare and liberty, which are associated with algorithmic ethics principles of beneficence, non-maleficence, autonomy, and explicability. Over-reliance

on fairness metrics would capture only one dimension of an algorithm's ethical impact.

As a step forward, our second contribution is the proposal of a generalised "Key Ethics Indicator" (KEI) approach that *explicitly* considers the ethical objectives, aligning to the contextual features that we have drawn out as important in ethical philosophy and welfare economics literature. The widespread discomfort with the use of ML to make decisions derives from the tension between the opportunity provided by algorithms that can more accurately predict an outcome and the risk of systematically reinforcing existing biases in the data and the risk of undermining human autonomy. On the other hand, unlike human subconscious biases, machine predictions can be systematically audited, debated, and improved. By understanding the holistic ethical considerations of each algorithmic decision-making process using KEIs, decision-makers can be better informed about the value judgements, assumptions, and consequences of their algorithmic design, opening up the conversations with regulators and with society on what is an ethical decision.

References

Nikita Aggarwal. 2018. Law and Autonomous Systems Series: Algorithmic Credit Scoring and the Regulation of Consumer Credit Markets. University of Oxford Business Law Blog (2018).

Nikita Aggarwal. 2020. The Norms of Algorithmic Credit Scoring. Available at SSRN 3569083 (2020).

Aristotle and TA Sinclair. 1962. Aristotle: The Politics; Translated with an Introduction by TA Sinclair. Penguin Books Limited.

Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. arXiv preprint arXiv:1909.03012 (2019).

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943 (2018).

Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 514–524.

Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A

study of bias in recidivism prediction instruments. Big data 5, 2 (2017), 153–163.

Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. 2021. Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 598–609.

Nancy S Cole. 1973. Bias in selection. Journal of educational measurement 10, 4 (1973), 237–255.

Patricia Hill Collins. 2002. Black feminist thought: Knowledge, consciousness, and the politics of empowerment. routledge.

Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023 (2018).

Kimberle Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. u. Chi. Legal f. (1989), 139.

Kevin P Donovan and Emma Park. 2019. Perpetual Debt in Silicon Savannah. Boston Review (2019).

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference. ACM, 214–226.

Ronald Dworkin. 1981. What is Equality? Part 1: Equality of Welfare. Philosophy and Public Affairs 10, 3 (1981), 185–246.

Margaret L Eaton. 2004. Ethics and the Business of Bioscience. Stanford University Press.

Ronel Elul and Piero Gottardi. 2015. Bankruptcy: Is It Enough to Forgive or Must We Also Forget? American Economic Journal: Microeconomics 7, 4 (2015), 294–338.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 259–268.

Marc Fleurbaey et al. 2008. Fairness, responsibility, and welfare. Oxford University Press.

Luciano Floridi and Josh Cowls. 2019. A unified framework of five principles for AI in society. Issue 1.1, Summer 2019 1, 1 (2019).

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im) possibility of fairness. arXiv preprint arXiv:1609.07236 (2016).

Pratik Gajane and Mykola Pechenizkiy. 2017. On formalizing fairness in prediction with machine learning. arXiv preprint arXiv:1710.03184 (2017).

David Gauthier. 1986. Morals by agreement. Oxford University Press on Demand.

Talia B Gillis. 2020. False Dreams of Algorithmic Fairness: The Case of Credit Pricing. Available at SSRN 3571266 (2020).

Barry Goldman and Russell Cropanzano. 2015. "Justice" and "fairness" are not the same thing. Journal of Organizational Behavior 36, 2 (2015), 313–318.

Jerald Greenberg. 1987. A taxonomy of organizational justice theories. Academy of Management review 12, 1 (1987), 9–22.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. CoRR abs/1610.02413 (2016). arXiv:1610.02413 http://arxiv.org/abs/1610.02413

Hoda Heidari, Michele Loi, Krishna P Gummadi, and Andreas Krause. 2018. A moral framework for understanding of fair ml through economic models of equality of opportunity. arXiv preprint arXiv:1809.03400 (2018).

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In Proceedings of the 2019 CHI conference on human factors in computing systems. 1–16.

Shelly Kagan. 1999. 30. Equality and Desert. What Do We Deserve?: A Reader on Justice and Desert (1999), 298.

Shelly Kagan. 2014. The geometry of desert. Oxford University Press.

Faisal Kamiran and Indrė Žliobaitė. 2013. Explainable and non-explainable discrimination in classification. In Discrimination and Privacy in the Information Society. Springer, 155–170.

Immanuel Kant and Mary Gregor. 1996. The metaphysics of morals. (1996).

Niki Kilbertus, Philip J Ball, Matt J Kusner, Adrian Weller, and Ricardo Silva. 2019. The sensitivity of counterfactual fairness to unmeasured confounding. arXiv preprint arXiv:1907.01040 (2019).

Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Fairness through computationally-bounded awareness. In Advances in Neural Information Processing Systems. 4842–4852.

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

James Rufus Koren. 2016. What does that Web search say about your credit? (Jul 2016). https://www.latimes.com/business/la-fi-zestfinance-baidu-20160715-snap-story.html

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. arXiv e-prints, Article arXiv:1703.06856 (March 2017), arXiv:1703.06856 pages. arXiv:1703.06856 [stat.ML]

Michelle Seng Ah Lee and Luciano Floridi. 2020. Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. Minds and Machines (2020). https://doi.org/10.1007/s11023-020-09529-4

Michelle Seng Ah Lee and Jatinder Singh. 2020. The Landscape and Gaps in Open Source Fairness Toolkits. Available at SSRN (2020).

Michelle Seng Ah Lee and Jatinder Singh. 2021. Risk identification questionnaire for unintended bias in machine learning development lifecycle. Available at SSRN (2021).

David Lewis. 1973. Causation. Journal of Philosophy 70, 17 (1973), 556–567.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. arXiv preprint arXiv:1803.04383 (2018).

Robert Nozick. 1974. Anarchy, state, and utopia. Vol. 5038. New York: Basic Books.

Derek Parfit. 1991. Equality or priority. University of Kansas, Department of Philosophy.

Robin A Prager et al. 2009. Determinants of the locations of payday lenders, pawnshops and check-cashing outlets. Federal Reserve Board Washington, DC.

John Rawls. 1999. A Theory of Justice, rev. ed.

John Rawls. 2001. Justice as fairness: A restatement. Harvard University Press.

Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. 2018. Aequitas: A Bias and Fairness Audit Toolkit. arXiv preprint arXiv:1811.05577 (2018).

Samuel Scheffler. 1994. The rejection of consequentialism: A philosophical investigation of the considerations underlying rival moral conceptions. Oxford University Press.

Amartya Kumar Sen. 1992. Inequality reexamined. Oxford University Press.

Andrew Van Dam. 2019. Searching for images of CEOs or managers? The results almost always show men. The Washington Post (3 01 2019).

Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In Proceedings of the 2018 chi conference on human factors in computing systems. 1–14.

Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In 2018 IEEE/ACM International Workshop on Software Fairness (FairWare). IEEE, 1–7.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2020. Why fair-

ness cannot be automated: Bridging the gap between EU non-discrimination law and AI. Available at SSRN (2020).

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law. Available at SSRN (2021).

Michael Walzer. 1983. Spheres of Justice. A Defense of Pluralism and Equality, New York, Basic (1983).

Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In CCF international conference on natural language processing and Chinese computing. Springer, 563–574.

Claire Zillman. 2019. Fortune 500 Female CEOs Reaches All-Time Record of 33. Fortune (16 05 2019).