Sahlgren, Otto, and Arto Laitinen. "Computing Apples and Oranges?: Implications of Incommensurability for (Fair) Machine Learning." (2022).

Computing Apples and Oranges? Implications of Incommensurability for (Fair) Machine Learning

Otto Sahlgren*[0000-0001-7789-2009] and Arto Laitinen*[0000-0002-4514-7298]

* Tampere University, Tampere, Finland otto.sahlgren@tuni.fi
arto.laitinen@tuni.fi

Abstract. In this paper, we discuss value relations and questions of incommensurability and incomparability in the context of machine learning and fairness therein. We examine three stances and consider their implications for machine learning supported decision-making and the pursuit of fair algorithms using a hypothetical example from recruitment.

Keywords: fair machine learning, philosophy, incommensurability, incomparability, value theory

Introduction

The value of machine learning (ML) systems lies in part in their capacity to discover patterns in data, to detect subtle (dis)similarities between data items, and to provide decision-makers information to help them make good decisions. Consider ML-based recruitment systems, for example, which allow for precise comparison and ranking of candidates in terms of their merits and overall "hiring-worthiness". However, to enable human decision-makers to make justified decisions, such systems should arguably track evaluative differences amongst candidates with respect to their "hiring-worthiness", which raises questions of comparability and commensurability. Recruitment can also have ethical objectives aside choosing the candidate with the most merits, for example, fairness. Similar questions arise here, as one should be able to compare competing ways to improve algorithms' fairness and to evaluate individuals' claims to fair treatment. For example, in Finnish Law, an individual belonging to an underrepresented group can be selected from among applicants that have "roughly the same qualifications", which raises the question of how "rough similarity" should be understood.

The present work is motivated by philosophical questions related to (in)comparability. The nature of value relations is significant in both cases described above – namely, in cases where we compare options with the help of ML systems and where we compare different ML systems in terms of ethical value. These topics are discussed in philosophical debates on (in)comparability and value (in)commensurability (see Chang, 1997; Elson, 2017; Hsieh, 2005) and recently also in the context of AI (see Dobbe et al., 2021; Fleisher, 2021; Goodman, 2021). We consider

the implications of three philosophical accounts of (in)commensurability and (in)comparability for the abovementioned cases.

The structure of the paper is as follows. In section 1, we review three philosophical accounts of (in)comparability and (in)commensurability and consider their implications for ML-supported decision-making. In Section 2, we consider further implications for building fairness-sensitive algorithms, a topic discussed under the umbrella of "fair ML". We distinguish three cases where incommensurability and incomparability pose both theoretical and practical challenges: conducting positive discrimination (or "affirmative action") with algorithms (section 2.1.), implementing individual fairness measures (section 2.2), and addressing trade-offs between statistical operationalizations of fairness (section 2.3). Throughout the paper, we use the context of ML-supported recruitment to illustrate the relevant questions and challenges.

1 Machine-Supported Decisions, Incomparability, and Incommensurability

The Trichotomy Thesis of value relations states that one of three relations – worseness, betterness, or equality – holds between any two items (e.g., options, value-bearers, and preferences) that are comparable. Debates on (in)commensurability and (in)comparability revolve around paradigmatic cases where we find ourselves amiss when determining which relation obtains, if any. For example, can Mozart and Michelangelo be compared in terms of creativity? If so, is the value relation in question captured by the trichotomy, or is there another relation?

The terms "incommensurability" and "incomparability" are used with slightly different meanings in the general debate (see Hsieh & Andresson, 2021). The seeming failure of comparison to yield any positive result in terms of betterness, worseness, or equality is called "incommensurability" by some and "incomparability" by others (and some use these terms interchangeably); we will not follow either of these usages. We will use the term "incommensurability" to refer to two things lacking a common measure (i.e., cardinal measuring is not possible) and "formal incomparability" to refer to two items lacking a covering value with respect to which they can be compared (e.g., the number 54 cannot be compared to the color green in terms of tastiness). We call the cases where two items are formally comparable, but neither is better, and yet they are not exactly equal, "puzzle cases".

We maintain that cases of formal incomparability are rather rare. In standard cases discussed in the literature, there is often a covering value (perhaps non-cardinal) that allows for formal comparability even in the puzzle cases (e.g., creativity in the case of Mozart and Michelangelo). Even though one can admit the covering value applies to both (Mozart and Michelangelo are arguably creative, while perhaps not exactly equally creative), the outcome of comparison remains puzzling.

We illustrate these questions with the following example:

> *A technology company wants to hire an ethicist. The chosen recruit should be "hiring-worthy" (i.e., possess a set of attributes, merits, and skills that make them a good technology ethicist and employee). "Hiring-worthiness" is the "covering value" here with respect to which the candidates are compared in multiple "component respects" (e.g., different skills and attributes). An initial screening shows that three roughly similarly merited candidates – Alex, Bill, and Connie – stand out from the crowd. Alex and Bill are academic philosophers with well-cited publications in technology ethics. Connie is a computer scientist with experience from ethical development in top companies in the industry. The recruiters decide to use a ML system to precisely rank the candidates. The system suggests that Alex is slightly better than Bill – perhaps, because Alex has published one paper more than Bill. Before the recruiters proceed to compare Alex and Connie, they ponder: is it even possible to rank the two? They have different degrees and backgrounds. Academia differs drastically from industry. Would they be comparing "apples and oranges"? Could the system, even in principle, help settle the choice between Alex and Connie?*

The previous hypothetical is analogous to Derek Parfit's (1987) famous example of comparing two poets and a novelist in terms of their literary merits which Parfit used to challenge the Trichotomy Thesis. In our case, if the Trichotomy Thesis is true, Alex's being equal to Connie and Bill's being worse than Alex would imply that Bill is worse than Connie (due to transitivity). But should one publication make the difference? The Thesis also implies that Alex and Connie are either equal or one is better. But how would we determine this in puzzling cases which involve apples and oranges, poets and novelists, or philosophers and computer scientists?

In the next three subsections we discuss three different responses to the puzzle cases. Some accounts deny the existence of such cases, some suggest they can be explained without rejecting the Trichotomy Thesis (Regan, 1977; Elson, 2017), and some grant the existence of a fourth relation, such as "parity" (Chang, 2002), "rough quality" (Parfit, 1987), or some type of "indeterminacy" or "imprecision". These are all loaded terms with their own connotations, and we will try to do justice to each while maintaining some terminological clarity. We will discuss the implications of each stance for ML-supported decision-making.

1.1 Eliminativism

It might seem that neither Alex or Connie is better than the other, or that nor are they equally "hiring-worthy". So-called *eliminativists* would maintain that this is a mere illusion. They argue that no two options objects are ever "apples and oranges" in a fundamental sense, and that the Trichotomy Thesis is true of any two items (Regan,

1997). Our epistemic limitations are the source of any apparent indeterminacy in comparison and ranking. The eliminativist would thus claim that the relations between Alex, Bill, and Connie in our example can all be accounted for with the three standard relations: each candidate is either better than, worse than, or exactly equal to another. For the eliminativist, the ML system can help determine which relations hold between the candidates provided that the relevant concepts ("hiring-worthiness") are clearly operationalized in the language of mathematics and that the model is good at tracking relevant the recruiters' preferences (e.g., the value contributed by publications). The eliminativist has no principled objection against ML systems' capacity to rank Alex and Connie *correctly* with respect to "hiring-worthiness" (even though existing systems might be limited in many contingent ways).

The eliminativist view has been criticized for many reasons, however. For one, epistemic limitations (e.g., ignorance) might not exhaust the problems with "puzzling cases". The parameters for choosing between two options are for the decision-maker to decide, the choice can yet be a "hard choice" (Chang, 2017). For example, one has firstperson authority in making the hard choice between a career in academic philosophy or in software development. Both are desirable in their own ways, yet they would still seem comparable in light of their contributions to a good life. Furthermore, while MLsupported decision-making is premised on the notion that value-bearers and preferences can in principle be mathematically represented and that human goals and tasks can be translated into computational problems with reward and loss functions, it might be that all "goals and purposes simply cannot be represented as the maximization of the expected value of a scalar reward" (Goodman, 2021, 5). Perhaps, "hiring-worthiness" does not lend itself to simple and exact quantification. ML systems would seem to provide a sense "quasi-precision" when dealing with such concepts and translating human goals into system specifications, respectively (see Dobbe et al., 2021).

1.2 "Parity"

Alternative accounts draw on the idea of there being "neighborhoods" of value and suggest that the Trichotomy Thesis does not capture the entire scope of value relations. Perhaps most influentially, Ruth Chang claims that "between two evaluatively very different items" (e.g., Alex and Connie, a poet and a novelist), "a small unidimensional difference cannot trigger incomparability where before there was comparability" (2002, 673). This claim is motivated with the so-called "Chaining Argument": Suppose there is another candidate Don, a philosopher similar to Alex and Bill but clearly worse than both in all component respects (e.g., publication, work experience, and so on). If Alex and Bill are in the same neighborhood of value as Connie, Don should plausibly be worse than Connie. If this is the case, however, Connie is comparable to Don and thereby also comparable to Alex and Bill because the philosopher candidates form a "chain", a sequence from worse to better philosophers. Hence Alex, Bill, and Connie are formally comparable. However, the kicker comes in Chang's suggestion that our

best philosopher, Alex, is not exactly equal to Connie. They are rather "on a par": formally comparable, not exactly equal, but neither is better than the other[1]. If they were exactly equal, "sweetening" Alex by making any small improvement (e.g., one more publication) would make sweetened "Alex+" better than Connie. When we compare Alex and Alex+, the latter is better. Nonetheless, Chang suggests that sweetening would not make a difference when it comes to Alex+ and Connie – they are not equal but Alex+ is not better. Both Alex and Alex+ are "on a par" with Connie.

Chang suggests "parity" deviates from the standard trichotomy of comparative relations. Whether it actually does is debated (see Hsieh & Anderson, 2021), but for present purposes we will we stay truthful to Chang's suggestion that it does. Now, if "parity" is a true comparative relation, ML systems should arguably be able to track such relations were they to estimate value relations, support human judgment, and provide justification for decisions. A challenge looms for existing ML methods, however, because parity is an intransitive relation: Alex is on a par with Connie, who is on a par with Alex+, but Alex is not on a par with Alex+ but is worse. The traditional trichotomy (betterness, worseness, and equality) is still there, however, because the fourth alternative is only added for puzzling cases (with apples and oranges). Regardless, allowing for "parity" in ML would require a way to implement this fourth value relation. For example, the output ranking would exhibit some type of a partial or perhaps incomplete ordering where a given rank (e.g., the top-*k* candidates) can contain items that are in a relation of "relaxed" or "imprecise" equality (because they are "on a par") as well as items that are exactly equal (because they are evaluatively identical).

1.3 "Clumpiness"

The idea of "parity" does not please all theorists as it requires accepting the intransivity of some value relations. In the ML setting, this can also prove challenging for precise ranking, for example. Is there a way, then, to accept the Trichotomy Thesis and thus avoid this possible issue without simultaneously falling into the eliminativist trap of "quasi-precision"? Hsieh (2005) suggests so, claiming that some values can be "clumpy". For example, the property of "hiring-worthiness" could be understood as a range property[2] within which we find "clumps" of "hiring-worthiness" similar to the clumps of "excellence" that different grades seek to track as evaluative categories for student coursework. For Hsieh, the number of clumps depends on the "resolution" of

[1] Derek Parfit (1987) has suggested similar view according to which the items can be "roughly equal". The difference between "rough equality" and "parity" is debated: according to one understanding, the former is "invoked to allow for comparability among alternatives that display the same respects" and the latter "to allow for comparability between alternatives that are different in the respects that they display" (Hsieh & Anderson, 2021, S2.2).

[2] A range property is a property which comes in degrees (e.g., a continuous scale along which to measure it) yet with respect to which we can specify a range and an object can either fall into that range (or not) in a binary sense. For example, an essay that is perfect in all evaluated respects falls perfectly within the range of essays worthy of the best grade. However, another essay that lacks slightly in certain respects can yet be equally worthy of the best grade.

comparison which "specifies the degree to which possession of each of the relevant respects of the covering consideration sorts an item into one clump or another" (2005, 184). Now, whereas Chang would suggest that items in a "clump" are perhaps rather

"on a par", Hsieh suggests they are in fact exactly equal. This has the happy consequence that the Trichotomy Thesis can be preserved without having to accept the eliminativist view that equal candidates must have the same real-valued "hiringworthiness". Due to the fixed resolution of comparison, the "top-notch" candidates, for example, are exactly equal in terms of hiring-worthiness. Yet this does not prevent one from scoring items (including candidates) on continuous scales nor "binning" them into clumps different ways. Different resolutions can serve practical aims: one can start with a coarser resolution of two clumps (e.g., "top-notch" and "unsuitable" candidates), and by increasing the resolution one can create new clumps (e.g., "moderately suitable").

Note that all comparisons need not be clumpy – there is room for totally ordered rankings and ML models which seek to track precise comparative value relations and generate corresponding rankings, respectively. Furthermore, even in cases where a target variable *is* supposed to track a clumpy value ("hiring-worthiness"), a continuous target variable provides one resolution of comparison, albeit a very fine-grained one. With clumpy values (e.g., hiring-worthiness), human decision-makers can build clusters, partial orderings, or output classes based on the real-valued total ranking of candidates, similar to how a teacher can determine which individuals should receive the best or worst grade after first observing and ranking a set of student essays. The important implication for ML models is that, when the target variable ought to track a clumpy value, the number and boundaries of the clusters, output classes, "bins", or ranks employed to evaluative items should reflect the appropriate resolution and outcome of (correct) evaluative judgment. The set of top-*k* candidates recommended to the recruiter, for example, should track the clump of "top-notch" candidates.

The key differences between eliminativism, and the "parity" and "clumpiness" views can thus be stated as follows: An eliminativist using the recruitment system would consider two candidates with different predicted values (e.g., Alex is 0.9 hiring-worthy, Bill is 0.89 hiring-worthy) as either better or worse than each other (as the ranked output would suggest). Were the eliminativist to have an ideal classifier, they could trust that the output ranking tracks the comparative value relations between the ranked items. A proponent of "parity" relations could consider Alex and Bill as unequal, whereas Alex would be on a par with Connie. Proponents of this view would thereby have to rely on partial order rankings. A proponent of clumpiness would require a similar approach. However, they would suggest that all candidates belonging to the same neighborhood of overall "hiring-worthiness" (e.g., the top-*k* candidates) should be viewed as exactly equal in the evaluative sense (despite the 0.1 difference between Alex and Bill). This is because two candidates having the same features (e.g., identical merits) is necessary and sufficient for exact equality in a descriptive sense, yet not necessary for exact equality in the evaluative sense (albeit sufficient). A ranking of candidates *qua* feature

vectors can be totally ordered, respectively, but an evaluative ranking of those candidates depends on the resolution which in turn specifies the clumps wherein candidates are evaluated as exactly equal. In other words, whereas "parity" would require a partially ordered ranking which allows for both equality and "parity" within a given rank, rankings with respect to clumpy values should have to allow for exact equality within a given rank in a way that does not imply numeric identity in real-valued outputs.

2 Fair Algorithms: Positive Discrimination, Similarity, and "Hard Choices" Concerning Fairness

The previous cases concerned comparability within the context of a single ML model seeking to track "hiring-worthiness". Often, however, ML-supported decision-making is guided also by ethical considerations related to fairness, for example. In this section, we consider three cases where questions of incommensurability and incomparability pose not only theoretical but practical challenges for designing algorithms in a fairnesssensitive manner. The first case concerns positive discrimination and its permissibility in (non-)automated recruitment. The second case concerns the measurement and implementation of individual fairness in ML systems. The third case concerns choices regarding trade-offs between multiple fairness targets in ML.

2.1 Imprecise Equality and Positive Discrimination with Algorithms

Recruitment policies can purposefully seek to promote disadvantaged groups' access to employment. One instrument for doing so is *positive discrimination* (or "positive action" or "affirmative action") where candidates from underrepresented protected groups (e.g., women) are favored over candidates from overrepresented groups (e.g., men). If applied with care, so-called "bias mitigation methods" developed for correcting discriminatory biases in ML systems (Mehrabi et al., 2021) could prove useful for such purposes in ML-supported recruitment as well. Importantly, however, justification of positive action often requires, among other things, that the selection process does not give a merely arbitrary or "disproportional advantage to members of the relevant group. In Finland, for example, the Non-Discrimination Act requires that all candidates are initially treated on an equal basis, and states that "an individual belonging to an underrepresented group can be selected from among applicants that have roughly the same qualifications" (Non-Discrimination Ombudsman of Finland, N.D.). As noted above, there is notable disagreement regarding how "rough sameness" should be interpreted. We will first consider what the formal setting of positive action implies for the use of bias mitigation methods after which we consider the notion of "rough sameness" through the theoretical lenses described in the previous section.

2.1.1. *Positive Action with Algorithms*

What kinds of bias mitigation methods would capture the "spirit of positive action" as described above? We suggest that at least three formal conditions need to be satisfied for a hiring decision to be considered "positive action" (in a "neutral" sense in which we can still ask whether it is justified):

- (*Formally Equal Assessment*): The selection process involves a formally equal assessment of a candidate from an underrepresented group (C_{UNDER}) and a candidate from an overrepresented group (C_{OVER}); their "hiring-worthiness" is assessed based on identical criteria. This implies the use of a single model with features representing component factors of "hiring-worthiness" and a single "cutoff" point (i.e., a decision-threshold).
- (*Absolute Sufficiency*): The chosen candidate is considered sufficiently hiringworthy in an independent, non-comparative sense. Depending on the case, the level of absolute sufficiency (i.e., the decision-threshold) can be decided either prior to or after ranking the relevant candidates.
- (*Imprecise Equality*): The favored candidate C_{UNDER} should be "roughly equal" in terms of their qualifications when compared to non-favored candidate C_{OVER}.

In other words, regardless of its justification, for recruitment to instantiate positive action at all, candidates C_{UNDER} and C_{OVER} should be "roughly equal" and "in the neighborhood" of what is required from each candidate in a non-comparative sense. Absolute Sufficiency is to be assessed with respect to a single covering value and by employing identical criteria that (hopefully) track the component factors of that value. For example, C_{UNDER} "making the cut" with lesser-than-sufficient qualifications would not qualify for positive action because it would not satisfy Absolute Sufficiency nor Imprecise Equality. Note that, in cases where Absolute Sufficiency has necessary conditions (call these *Criteria-First Cases*), a candidate that satisfies those conditions is plausibly "clearly better" than one that does not. In other cases, such as when the level of absolute sufficiency can be set only after ranking the candidates, the best candidate takes the spot (call these *Ranking-First Cases*). In Aggregation Cases, C_{UNDER} and C_{OVER} need not be on the same side of the decision-threshold, however, even though the recruiter cannot choose C_{UNDER} if C_{OVER} is somehow "clearly better".

Positive action involves taking protected attributes into account in selection after the initial assessment has taken place. Consider now, that there are at least three general different ways to incorporate information about protected group-membership into MLsupported decision-making (see Hellman, 2019; Mehrabi et al., 2021). First, one might use different decision-thresholds for members of different protected groups within the model. Second, one might use different models for those groups entirely. Third, one could change the output labels of members of the underrepresented group from negative to positive when they are close to the decision-threshold (Kamiran et al., 2012).

We suggest that none of these methods satisfy all three conditions described above. Consider the first approach. Here, recruiters would employ similar component factors for comparison but with different "cut-off points" for C_{UNDER} and C_{OVER}. The level of Absolute Sufficiency would thereby differ across groups, implying that different weights are given to component factors. If so, the condition of Formally Equal Assessment is violated. In the second approach, "hiring-worthiness" would be evaluated with different models for C_{UNDER} and C_{OVER}. This means the covering values used for evaluation are non-identical and the Formally Equal Assessment condition is again violated[3]. The third approach comes closest the formal setting assumed in cases of positive action: candidates are evaluated based on the same model and there is a

single decision-threshold. However, it lacks a way to resolve cases where there are two candidates – C_{UNDER} and C_{OVER} – and a single available position. Positive action will be at best an artefact of choosing C_{UNDER} as opposed to an explicit aim implemented in the method's processing logic in a strict sense. That is, the comparative dimension of positive action – "if two candidates are within a range *R*, choose C_{UNDER} over C_{OVER}" – remains uncaptured by the method.

2.1.2. *Positive Discrimination and "Rough Sameness"*

Let us now consider Ranking-First Cases to examine how theoretical stances concerning comparative relations bear on (the possibility of) positive action. First note that eliminativists and other proponents of the Trichotomy Thesis often consider a choice permissible only if there are normative reasons to choose A over B (e.g., A is better) or if A and B are exactly equal. For them, the boundaries of "rough sameness" with respect to qualifications will remain rather arbitrary and thus they would consider it irrational to choose C_{UNDER} over C_{OVER} in case the former is ranked below the latter. The eliminativist can, of course, concede the plausible claim that there are other normative reasons (e.g., the value of equity) that override rational choice based on qualifications alone.

A proponent of clumpy values could contend, however, that "roughly the same qualifications" merely means that C_{UNDER} and C_{OVER} have to belong to the same clump of "hiring-worthiness". As C_{UNDER} and C_{OVER} would hence be exactly equal, choosing either one of them is permissible in light of rational choice based on qualifications alone (contra the eliminativist) and the instantiation of positive action is merely an artefact of choosing C_{UNDER}. However, one could question whether "rough sameness" should actually be understood as an even softer requirement: could it not suffice that C_{UNDER} belongs merely to the clump below the one including C_{OVER}?

Alternatively, if "rough sameness" is what Chang means by "parity", positive action is possible when C_{UNDER} and C_{OVER} are "on a par" – neither candidate is actually better than the other, nor are they exactly equal. Both "parity" and "clumpiness" can thereby

[3] An open question is, however, whether the non-identical models can be commensurate as models of overall "hiring-worthiness".

lead to the conclusion that one never hires a *de facto* worse candidate when implementing positive action.[4] While we do not discuss these issues further, we note that each stance has implications for how positive action ought to be understood as a non-moralized category of acts, and how it should be implemented in ML-supported decision-making.

2.2 Tracking Evaluative Features: The Case of Individual Fairness

Individual fairness (IF) as an approach to algorithmic fairness draws on the principle of formal equality: fairness requires treating similar individuals similarly (Dwork et al. 2012). Individual fairness is measured with similarity-metrics that estimate the similarity (or distance) between individuals in terms of some set of attributes which typically exclude an individual's protected status, for example. If individuals who are similar according to the metric receive different outputs, the algorithm is considered unfair.

Similarity metrics ought to track moral values and (dis)similarities which are relevant from the perspective of fairness. Will Fleisher (2021) considers incommensurability to pose a challenge because "a similarity metric requires that it be possible to aggregate the moral values, or evaluate them together, in a straightforward way" (2021, 17). Fleisher notes that some "moral values are incommensurable" and "cannot be evaluated on a common measure, i.e., they cannot be straightforwardly aggregated or exchanged" (Ibid., 3). Indeed, "similarity" in IF approaches is ambiguous between exact, "descriptive" similarity (being qualitatively identical) and similarity that is relevant from the perspective of fairness and ethics. However, the relationship between these two types of similarities can be complex – a (dis)similarity of the former kind may or may not equate to (dis)similarity of the latter kind. As we considered above, it might be that "sweetening" one candidate would not render them better than another candidate, even though the descriptive distance between them would increase as a result (see Fleisher, 2021). Likewise, if values are clumpy, for example, a small difference between two "top-notch" candidates will not make a relevant difference in terms of fairness.

Estimating "fairness-relevant" distance and possible puzzles in how evaluative features behave (e.g., "parity" and "clumpiness") requires human value judgements.

[4] If parity obtains between two options, one will lack so-called "given reasons" (i.e., reasons grounded in normative facts) for choosing between them. One has no *prima facie* reason to favor either, making the decision a "hard choice" (Chang, 2017). Chang argues that normative commitments can create "will-based reasons" that function as tie-breakers. Positive action could thus be understood as a normative commitment to equality that creates a reason for choosing between candidates that are "on a par". *Mutatis mutandis*, the same holds for the clumpiness view.

Fleisher notes, however, that appealing to human arbiters in evaluating similarities between individuals, while promising, can be problematic due to "implicit biases in [human] judgment" (2021, 2). Fleisher is correct, but the cure may nonetheless be adding more human arbiters, and the hope that different humans have different biases. In most cases, cognitive and socio-cultural diversity among those arbiters can be beneficial to ensure "diversity of biases" and that the arbiters arrive at correct judgments regarding how similarity ought to be measured. Even if we accept that measurements of similarity will always reflect prior moral judgments and biases (Fleisher, 2021, 2), it is fine (and inevitable) that judgements about fairness rely on evaluative or moral judgements insofar as such an equilibrium is achieved.[5]

2.3 "The Impossibility of Fairness" and "Hard Choices"

A final challenge relates to trade-offs in implementing multiple operationalizations of algorithmic fairness. As Dobbe and colleagues note, "normative concerns of comparable significance and scope must be rendered commensurable in order for a

responsible tradeoff to be struck and translated to a system's specification" (2021, 4). However, trade-offs between different fairness criteria are a prime example of cases where we seem to be lacking such commensurate options. Various fairness definitions prescribe equalizing some statistical metric (e.g., positive predictions, error rates) across individuals or groups in the model (Verma & Rubin, 2018). But many of these metrics cannot be equalized simultaneously except in highly contrived cases (Kleinberg et al., 2017). Trade-offs are taken to suggest "the impossibility of fairness" due to fairness definitions' representing irreconcilable moral views.

To make justifiable decisions concerning trade-offs, we should be sensitive to reasons for choosing one fairness metric over another. However, some consider the decision regarding proper measurement and implementation of fairness "a hard choice" in virtue of the options being incommensurate: "certain alternatives are neither better, worse nor equal to one another with respect to fairness" (Goodman, 2021, 7). Rival options are rather "on a par". Value incommensurability thus poses a challenge for justifiably deciding which fairness criteria to implement in algorithms. But if such a decision requires choosing between incommensurate options, what do we do?

Goodman's proposal draws on Chang's (2017) view according to which "will-based reasons" created through normative commitment can function as tie-breakers:

"[n]othing in the world will tell us the correct answer" regarding proper measures of algorithmic fairness; "[i]nstead, we must *commit*" (Goodman, 2021, 7). Dobbe and colleagues (2021) propose another solution. Recognizing that ML systems typically affect numerous stakeholders with different interests, they suggest that incommensurability and hard choices become *political* issues as AI systems' normative capacities cannot be evaluated and measured by the same standards by different

[5] Fleisher also considers a possibility we already discussed above: using partial order rankings instead of (or in addition to) similarity-metrics.

stakeholders. We note that these solutions are not mutually exclusive – whether it is individual persons or collectives making significant decisions, the cold reality of compromises is equally faced by both. Extending Chang's solution to "hard choices" between incommensurable options, one could argue that collective wills can create tiebreaking "will-based reasons" through normative commitment in a manner analogous to the individual case.

If trade-offs are exclusively "hard choices", there are no normative facts or secondorder principles concerning distributive fairness that provide reasons to resolve tradeoffs in a manner *R* purely because *R*-ing is the right thing to do. We should resort to proceduralism (e.g., fair democratic decision-making procedures) or will-based reasons (e.g., resolution through normative commitment), for example. While these might be independently justifiable and desirable approaches, we note that equating trade-offs with Changian "hard choices" denies the possibility of there being "given" normative reasons that speak in favor of resolving trade-offs in one way over another. For instance, while a fair democratic process might lead to the decision to choose the option that creates the largest benefit to those who are worst off, one could also argue that prioritarian regard should guide choices concerning trade-offs because it is what fairness requires.

Furthermore, it is noted in some works that certain fairness definitions are in principle compatible albeit practically in conflict due to contingent states-of-affairs (e.g., differences in the distribution of attributes across subpopulations) (Binns, 2020). Some trade-offs can thus arise "merely" in virtue of our present, contingent circumstances. This suggests, however, that they can be resolved in the long-term, at least in principle. We would need a "transitional approach" and incremental improvements to get to a place where multiple fairness definitions can be simultaneously satisfied. If this is true, it leaves open the possibility that different fairness metrics for algorithms may in theory be commensurate with respect to a covering value: Fairness or Justice *qua* an ethico-political value. Different operationalizations of fairness in ML could be understood as comparable *qua* factors that contribute to a multidimensional covering value understood as overall fairness (or not). Competing fairness measures are applicants for the job of the best conception of Fairness (or at least one component of it), as it were. As we are not ideal judges with direct access to ideals of Justice and Fairness, we do not merely choose or commit to any of the candidates. We test them, build new ones, compare their pros and cons, and hopefully end up with the best conception of algorithmic fairness so far. Ultimately, the right approach to resolving trade-offs will be one that best reflects the constitutive aim of conceptions of algorithmic fairness; what "Fairness with a capital F" or "Justice with a capital J" in fact requires. Insofar as feasibility constraints and long-term effects might prevent us from achieving the best solution straight away, for the time being, one might have to settle for the second-best.

3 Conclusions

In this paper, we sought to motivate questions related to incommensurability, incomparability, and "hard choices" in the context of ML-supported decision-making. We reviewed three stances on comparability and value relations, discussing their implication for ranking and ordering items with ML. Each stance, we suggested, has implications concerning whether and how ML-generated rankings can track value relations. We also discussed fairness in ML, noting how similar puzzles arise in the context of building fair algorithms. Here, we located challenges relating to positive discrimination and how it might be pursued with algorithms and for determining whether and how to implement one or several fairness measures in ML systems. While the fundamental questions concerning the nature of value relations remain unsolved, we hope to have shed light on their practical significance for designing and using ML systems, suggesting also possible ways forward.

References

Binns, R. (2020). On the apparent conflict between individual and group fairness. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 514-524.

Chang, R. (1997). Introduction. *Incommensurability, Incomparability, and Practical Reason*, R. Chang (ed.), Cambridge: Harvard University Press.

(2002). The Possibility of Parity. *Ethics*, 112, 659–688.

(2005). Parity, Interval Value, and Choice. *Ethics*, 115, 331–350.

(2017). Hard choices. Journal of the American Philosophical Association, 3(1), 1-21.

Dobbe, R., Gilbert, T. K., & Mintz, Y. (2021). Hard choices in artificial intelligence. Artificial Intelligence, 300, 103555. https://doi.org/10.1016/j.artint.2021.103555.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214-226.

Elson, L. (2017). Incommensurability as Vagueness: A Burden-Shifting Argument. *Theoria*, 83 (4), pp. 341–363.

Fleisher, W. (2021). What's Fair about Individual Fairness?. *AIES '21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 480–490.

Goodman, B. (2021). Hard Choices and Hard Limits for Artificial Intelligence. *arXiv preprint arXiv:2105.07852.*

Hellman, D. (2020). Measuring algorithmic fairness. Virginia Law Review, 106(4), 811-866.

Hsieh, N. H. (2005). Equality, clumpiness and incomparability. Utilitas, 17(2), 180-204.

Hsieh, N-H. & Andersson, H. (2021). Incommensurable Values. The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.). https://plato.stanford.edu/archives/fall2021/entries/value-incommensurable/. Accessed 10.5.2022.

Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. *2012 IEEE 12th International Conference on Data Mining*, 924-929. IEEE.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.

Non-Discrimination Ombudsman of Finland. (N.D). *Positive action*. [Website]. https://syrjinta.fi/en/positive-action. Accessed 9.5.2022.

Parfit, D. (1987). *Reasons and Persons*. Oxford: Oxford University Press.

Regan, D. (1997). Value, Comparability, and Choice. *Incommensurability, Incomparability and Practical Reason*, In R. Chang (ed.). Cambridge: Harvard University Press.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), 1-7. IEEE.