Algorithmic Fairness as an Inconsistent Concept

Unpublished work, final author version (accepted manuscript after peer review),

forthcoming in: *American Philosophical Quarterly*, 62.1.

Algorithmic Fairness as an Inconsistent Concept

Patrik Hummel

Abstract: In this article, I investigate whether *algorithmic fairness* is an inconsistent concept (the inconsistency thesis). Drawing on the work of Kevin Scharp, inconsistent concepts can apply and disapply at the same time (2.). It is shown that paradigmatic issues of algorithmic fairness fit this description (3.). Similarities and differences to received views (4.) and alternatives to the inconsistency thesis are considered (5.). Suggestions are articulated on how the inconsistency thesis might hold ground nevertheless, or at the very least denotes a distinctive option in argumentative space whose status and implications merit further evaluation.

Keywords: Fairness, Algorithms, Inconsistency, Impossibility, Concepts

1. INTRODUCTION

The discourse on algorithmic fairness investigates under which conditions classifications provided by algorithms are fair or unfair (Barocas, Hardt, and Narayanan 2019; Hutchinson and Mitchell 2019; Mitchell et al. 2021). This is a fundamentally important project in view of the disruptive and transformative effects of state-of-the-art algorithms (Burrell and Fourcade 2021; Hopster 2021; Löhr 2023). One widely discussed difficulty is that different understandings and formalizations of algorithmic fairness pull in different directions (Chouldechova 2017; Corbett-Davies et al. 2017; Kleinberg, Mullainathan, and Raghavan 2017; Hutchinson and Mitchell 2019). While this finding has stimulated further work on how to pursue algorithmic fairness in practice (Whittaker et al. 2018; Whittlestone et al. 2019), little attention is being devoted to the implications of widely discussed incompatibilities and tradeoffs for the concept of algorithmic fairness

itself. The present article articulates a proposal on this neglected topic: drawing on Kevin Scharp's work on inconsistent concepts (2.), it formulates and investigates what I call the inconsistency thesis: the view that various findings about the relation between different fairness considerations—specifically separation and sufficiency, calibration and redlining, and so-called 'costs' of fairness—suggest that algorithmic fairness is an inconsistent concept (3.). The main goal of the paper is to suggest that the inconsistency thesis is a distinctive option in argumentative space, an option which so far we have neither anticipated, nor ruled out, nor even clearly articulated. Similarities and differences between the inconsistency thesis and received views (4.) and alternatives to the inconsistency thesis are considered (5.). Throughout, it is assumed that the inconsistency thesis is relevant to the pursuit of algorithmic fairness, especially insofar as inconsistency leads to defective thoughts. The present contribution, however, is focused on a specification of the inconsistency thesis, consideration of salient criticisms of the thesis, and tentative responses to these criticisms. Practical implications of the inconsistency thesis (if true) are beyond the scope of this text and left for a different occasion. Besides shedding light on the concept of algorithmic fairness, this is a case study of what the hypothesis or diagnosis of inconsistency implies in a field of technology that disrupts and reconfigures social reality.

2. INCONSISTENT CONCEPTS

Concepts play an important role in philosophy, psychology, and cognitive science. They are seen as the "building blocks of thought" (Margolis and Laurence 2022). Concepts come in different shapes and serve different purposes. Deploying a concept is typically a way of trying to partition, make sense of, and navigate reality. Beyond this rough gloss, it is challenging to state what a concept is in theory-neutral terms. For example, foundational disagreements revolve around whether concepts are mental representations in the minds of thinkers, abilities of thinkers, abstract objects existing independently of the minds of thinkers, or whether they are something else entirely. In the following, these differences will be sidestepped as far as possible and we will stay neutral on the exact nature of concepts.

Kevin Scharp (2013) argues that some concepts are inconsistent. Philosophers operate with various formulations of what consistency and inconsistency involve. For example, a consistent set of statements is

one from which you cannot prove a contradiction (Sider 2010, 62). An axiomatic system is consistent if and only if not every well-formed formula is a theorem of that system; a system is inconsistent if and only if every well-formed formula is a theorem (Hughes and Cresswell 1996, 46). Not much hinges on the particular formulation; logicians regard these standard definitions as equivalent.

In Scharp's terminology, concepts are individuated by what he calls constitutive principles. Constitutive principles are rules that specify if and when a concept applies and/or disapplies. For example, the concept *bachelor* (henceforth, italics are used to indicate reference to a concept) has as its constitutive principles that it applies to individuals who are male and unmarried, and disapplies to individuals who are female or married. These rules are constitutive of the concept in the sense that they individuate that concept and state what falling under the concept requires. Constitutive principles "tell you which concept it is" (Scharp 2020).

Inconsistent concepts as understood by Scharp have constitutive principles of a particular kind. Roughly, "a concept is inconsistent iff its constitutive principles are inconsistent" (Scharp 2013, 36). This is the case if at least in some situation, application of one principle necessitates the disapplication of another, or vice versa. He discusses several examples of inconsistent concepts. For instance, he defines the concept *rable*, which has two constitutive principles:

(1a) *rable* applies to x if x is a table.

(1b) *rable* disapplies to x if x is a red thing (Scharp 2013, 36).

While these two principles are not logically inconsistent, a contradiction results once one is confronted with a red table. In such a case, *rable* applies and disapplies at the same time. On this basis, one would judge that the entity is a *rable* and that it is not the case that that entity is a *rable*. This is a problem insofar as "most people do not believe that any contradictions are true (even ones involving odd concepts like rable)" (Scharp 2013, 36). Besides being led into contradiction, inconsistent concepts can lead thinkers to subscribe to false empirical claims: "Assume for reductio that some red tables exist. Let R be a red table. The reasoning above shows that R is a *rable* and R is not a *rable*. Contradiction. Therefore, no red tables exist. We have proven an obviously false empirical sentence using only logic and the constitutive principles for 'rable'" (Scharp

2013, 36). Scharp concludes that "adding rable to one's conceptual repertoire corrupts it in a certain way" (Scharp 2013, 36).

Scharp's seminal and most extensive case study of an inconsistent concept concerns *truth* (Scharp 2007; 2013). In so-called liar paradoxes, such as 'This sentence is false,' the concept applies and disapplies at the same time. He proposes to replace the concept of *truth* with two different concepts, *ascending-truth* and *descending-truth*. In fact, Scharp's view is that not just the study of *truth*, but philosophy in general "is for the most part the study of inconsistent concepts" (Scharp 2020, 397): for topics such as knowledge, nature, meaning, virtue, explanation, and many others, what philosophical critique brings out is that the respective concept under consideration is inconsistent. The relevant positive, constructive projects philosophers are engaged in are most plausibly construed as attempts to replace the concept found to be inconsistent with one or more ameliorated (Haslanger 2000; 2012) concepts.

In view of examples like these, Scharp argues that inconsistent concepts are defective. Inconsistent concepts lead thinkers into contradiction in a world that, as many (though admittedly not all (Priest, Berto, and Weber 2022)) argue, does not contain true contradictions. Such concepts can thus lure thinkers into commitments to false claims. As such, they are clumsy means to achieve our ends—even if in everyday circumstances, the defects of inconsistent concepts like *truth* need not be immediately apparent.

3. ALGORITHMIC FAIRNESS AS AN INCONSISTENT CONCEPT

Why would we think that the notion of an inconsistent concept is relevant to algorithmic fairness? It is because *algorithmic fairness* instantiates the distinctive features of inconsistent concepts. The present chapter (3.) will state and provide support for the inconsistency thesis by reference to a selection of exemplary issues, though with no claim to exhaustiveness. I shall assume that there is a concept of *algorithmic fairness*, and that at least a subset of its constitutive principles can be specified.

3.1 Separation and Sufficiency

As one motivation for the inconsistency thesis, consider the widely discussed case of the Correctional Offender Management Profiling for Alternative Sanction (COMPAS) algorithm, which commentators have

interpreted as a standoff between *separation* and *sufficiency* (Hellman 2020). Barocas et al. (2019, 45–75) formalize these two criteria as follows: for an algorithmic classifier or score R, attribute A, and target variable Y (and with $\perp$ denoting statistical independence),

- random variables (R,A,Y) satisfy *separation* if $R \perp A \mid Y$;
- random variables (R,A,Y) satisfy *sufficiency* if $Y \perp A \mid R$.

Each of these criteria can be given an intuitive motivation. Hellman comments: "one measure [s*ufficiency*] begins with the score and asks about its ability to predict reality. The other measure [*separation*] begins with reality and asks about its likelihood of being captured by the score" (2020, 816).

For both notions, there are closely related, relaxed or equivalent criteria such as *balance* (Kleinberg, Mullainathan, and Raghavan 2017, 5) for *separation*, and *calibration* (Barocas, Hardt, and Narayanan 2019, 51) for *sufficiency*. For example, *calibration* interprets the risk scores as probabilities. The criterion is satisfied if "the set of all instances assigned a score value r has an r fraction of positive instances among them". Applied across groups, c*alibration* requires that for all score values r and groups a, Probability $\Pr\{Y = 1 \mid R = r, A = a\} = r$ (Barocas, Hardt, and Narayanan 2019, 51).

With the content of s*eparation* and *sufficiency* at hand, let us turn to COMPAS. This algorithm provides risk scores about defendants, including for "pretrial release", "general recidivism", and "violent recidivism" (Northpointe 2015, 26–31). Concerns about the fairness of these scores were prompted by a report and accompanying data set published by ProPublica (Angwin et al. 2016; Larson et al. 2016). Besides raising more general worries about the accuracy of COMPAS (which parts of ProPublica's analysis suggested to be only "somewhat more accurate than a coin flip" (Angwin et al. 2016)), the report was particularly concerned with an apparent difference in the treatment of Black and White defendants: "In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways. The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants. White defendants were mislabeled as low risk more often than black defendants" (Angwin et al. 2016). The worry thus is that

COMPAS fails to satisfy s*eparation* (Hellman 2020, 816): conditional on the defendant not actually recidivating (Y), Black defendants (A) do receive a substantially higher risk score (*R*).

In response, COMPAS producer Northpointe (now called Equivant) responded by "strongly reject[ing] the conclusion that the COMPAS risk scales are racially biased against blacks" (Dieterich, Mendoza, and Brennan 2016). Its defense was primarily based on the assertion that COMPAS does satisfy s*ufficiency* (Hellman 2020, 816), given that it "can discriminate recidivists and non-recidivists equally well for two different groups such as blacks and whites" (Dieterich, Mendoza, and Brennan 2016, 9) and that "the probability of recidivating, given a high risk score, is similar for blacks and whites" (Dieterich, Mendoza, and Brennan 2016, 9). Conditional on a particular risk score (*R*), there is statistical independence between actually recidivating (*Y*) and being a Black defendant (*A*).

Regardless of where one stands in this debate, each discussant arguably puts forward a legitimate concern. ProPublica is right to complain about the fact that Black individuals who do not recidivate receive higher COMPAS risk scores than non-Black individuals who do not recidivate. At the same time, Northpointe does have a point in highlighting that a particular COMPAS risk score is (let us assume) equally accurate, regardless of whether the defendant is Black or White. Taken at face value, we might thus conclude from the COMPAS controversy that *algorithmic fairness* has at least two constitutive principles:

(2a) *algorithmic fairness* applies if *sufficiency* is satisfied;

(2b) *algorithmic fairness* disapplies if s*eparation* is undercut.

Even if deemed incomplete and providing only broad brushstrokes, an analysis of *algorithmic fairness* in terms of (2a) and (2b) suggests that the concept applies and disapplies simultaneously to COMPAS. It is thus an inconsistent concept in Scharp's sense. COMPAS is of course not an isolated case: as already anticipated early discourses on fairness in classification and prediction in the 1970s (Hutchinson and Mitchell 2019, 50–51) and proven more recently by several authors (Chouldechova 2017; Kleinberg, Mullainathan, and Raghavan 2017; Berk et al. 2018), setting aside highly contrived scenarios, it is practically impossible to satisfy (2a) and (2b) simultaneously. Demonstrations of such incompatibility are

widely called 'impossibility theorems' about *algorithmic fairness*. As Kearns and Roth put it, we find "fairness fighting fairness" (2020, 84).

3.2 Calibration and Redlining

There are further routes to motivate the inconsistency thesis. Suppose we subscribe to constitutive principle (2a): *algorithmic fairness* applies if *sufficiency* is satisfied, in particular if c*alibration* holds. Corbett-Davies and colleagues note that this condition is compatible with discriminatory practices of the following kind (Corbett-Davies et al. 2017, 803–4). In their example, a 30% detention risk threshold is used for all defendants. There are two groups, each with an actual 20% reoffending rate. However, there is a vicious, deliberate asymmetry in the data about both groups: one of them is segmented into low (≥10% chance of reoffending), average (≥20%), and high risk (≥40%), whereas all members of the other group are classified as average risk (≥20%). As a result, none of the latter are detained. Still, "this is algorithm is calibrated" as anyone "labeled average risk reoffend[s] 20% of the time" (Corbett-Davies et al. 2017, 804), that is, the risk score means the same for both groups (see also Eva 2022, 253–55). A similar principle is followed in the historical practice of redlining, though by "ignoring information about the disfavored group" (Corbett-Davies et al. 2017, 804), for example when basing decision-making about financial services partly on whether an applicant resided in a neighborhood with low average income. By the same reasoning, such an algorithm can satisfy *calibration*, but in doing so lumps together all members of the respective low-income-area residents into one risk cluster, whereas members of the favored group enjoy a more fine-grained stratification, with some of them receiving the loan.

In both the hypothetical and the historical case, *sufficiency* or *calibration* is achieved, but the granularity of information about one group is deliberately modified in ways that unfairly disadvantage one group. Once again, we are led into contradiction: (2a) is satisfied, yet:

(2c) *algorithmic fairness* disapplies in redlining cases, that is, cases in which differences in the granularity of considered information about different groups lead to (un)favorable treatment of one or more groups.

3.3 Costs of Fairness

Yet another route to inconsistency results if one broadens the focus beyond continuous risk scores as in the aforementioned examples, and towards binary decision rules on how to proceed in view of such risk scores. Corbett-Davies and colleagues (Corbett-Davies et al. 2017; Corbett-Davies and Goel 2018) argue that one consideration with a COMPAS-style algorithm is to maximize public safety, which requires "detaining all individuals deemed sufficiently likely to commit a violent crime" (2017, 804), that is, those assumed to exceed a particular risk threshold. The level of this threshold will be a function of the number of individuals society is willing to detain in order to prevent one additional violent crime. Corbett-Davies and colleagues go on to demonstrate the following insights for at least three fairness criteria: statistical parity, conditional statistical parity, and predictive equality (a relaxation of s*eparation*). For the sake of the argument, suppose that, maybe in view of ProPublica's arguments, one accepts the following constitutive principle:

(2d) *algorithmic fairness* applies if s*eparation* is satisfied.

3.3.1 Group-specific thresholds

Corbett-Davies and colleagues show that in a COMPAS-style scenario, optimal algorithms satisfying any of the surveyed fairness criteria, including *separation,* entail group-specific risk thresholds for detention. For example, they require detaining a member of one group if the probability of recidivism is deemed to exceed 20%, and detaining a member of a different group if the probability of recidivism is deemed to exceed 10% (Corbett-Davies et al. 2017, 799). In other words, there is a "need for race-specific decision thresholds to achieve prevailing notions of algorithmic fairness" (Corbett-Davies et al. 2017, 801). Such group-specific thresholds, while needed to satisfy the reviewed fairness criteria, simultaneously undermine the case for *algorithmic fairness*. Intuitively, it is unfair if members of one group are detained at a given level of estimated risk, while members of a different group are released at the very same level of estimated risk. In other words, Corbett-Davies et al. demonstrate "an inherent tension between satisfying common fairness constraints and treating all individuals equally, irrespective of race" (2017, 801). The former desideratum forces us to undercut the latter. Indeed, Corbett-Davies et al. assume that such thresholds could

trigger *strict scrutiny* under the Equal Protection Clause of the Fourteenth amendment (2017, 804; Hellman 2020, 852–53). This suggests that

(2e) *algorithmic fairness* disapplies if classification is based on group-specific risk thresholds, that is, individuals are held to different standards.

As per (2d) and (2e), *algorithmic fairness* can apply and disapply simultaneously because group-specific risk thresholds are compatible with (in fact as Corbett-Davies et al. argue: necessary for) the satisfaction of *separation*.

3.3.2 Overly strict and/or overly lenient thresholds

As an alternative to group-specific thresholds, Corbett-Davies and colleagues consider an unconstrained algorithm which "requires applying a single, uniform threshold to all defendants" (2017, 797). This single-threshold algorithm "maximizes public safety while also satisfying one important understanding of equality: that all individuals are held to the same standard, irrespective of race" (2017, 797). It does not, however, satisfy any of the reviewed fairness criteria, which would operate as constraints on that optimal single-threshold algorithm. This marks a principled "tension between improving public safety and satisfying prevailing notions of algorithmic fairness," effectively raising a "cost of fairness" (2017, 797). In fact, they show by reference to ProPublica's dataset on COMPAS that the cost of satisfying the reviewed fairness criteria is a cost not only to public safety by way of increases in violent crime; there is also a cost by way of increases in the detainment of low-risk individuals for whom the unconstrained algorithm would recommend release (2017, 802).

Corbett-Davies & Goel (2018) put the point by reference to what they call classification parity. This family of fairness criteria demands "that some given measure of classification error is equal across groups" (Corbett-Davies and Goel 2018, 5). *Separation*, which requires parity in error rates, is an instance of classification parity. Corbett-Davies & Goel caution that relative to the optimal single risk threshold for all groups, the group-specific thresholds needed to satisfy any of the criteria subsumed under classification parity have adverse effects on group well-being. In their example, they assume that the optimal recidivism risk threshold for detention is 25%, and that equalizing the false positive rates of two groups would imply a

16% threshold for one group and a 31% threshold for the other. In other words, the thresholds needed to satisfy classification parity will be "overly strict" for one group and "overly lenient" for another (Corbett-Davies and Goel 2018, 14). The thresholds are "overly" strict (or lenient) and thereby unfair in the sense that they require more (or less) detention than the singular threshold flowing from society's willingness to detain in order to prevent additional crime (for related worries, see Holm 2023). Once again, we are led into contradiction:

(2f) *algorithmic fairness* disapplies if one group faces an "overly strict" or "overly lenient" (Corbett-Davies and Goel 2018, 14) detention thresholds.

As per (2d) and (2f), *algorithmic fairness* can apply and disapply simultaneously because overly strict or lenient thresholds are compatible with (in fact as Corbett-Davies et al. argue: necessary for) the satisfaction of s*eparation*.

4. THE CONTOURS OF THE INCONSISTENCY THESIS

In view of these motivations of the inconsistency thesis, it is worth highlighting a selection of similarities and differences between the thesis and various received views.

To begin with, s*eparation*, *sufficiency*, and other notions are often called fairness "metrics" (e.g., Binns 2018; Mitchell et al. 2021; Bell et al. 2023), "measures" (e.g., Hellman 2020; Defrance and De Bie 2023), "criteria" (e.g., Corbett-Davies and Goel 2018; Hutchinson and Mitchell 2019; Hedden 2021), or "definitions" (e.g., Berk et al. 2018; Verma and Rubin 2018; Saxena et al. 2019; Mehrabi et al. 2021; Tsamados et al. 2022). Some authors move back and forth between these terms (e.g., Kearns and Roth 2020). It is rarely noted (though see Fleisher (2021)) that the first three phrasings allow for a purely epistemic reading on which they stay agnostic on what fairness consists in, and express proposals solely on how to discern and to measure fairness—just as we might discern a car's speed by checking the speedometer, without thereby supposing that the state of its speedometer is what its speed consists in. While this distinction deserves more attention than it has received in connection with *algorithmic fairness*, the inconsistency thesis aligns with authors speaking of fairness 'definitions': it considers *separation*, *sufficiency*, and other notions not just as evidence for, but as constitutive of *algorithmic fairness*.

This being said, the analyses just discussed are plausibly just partial. For example, neither ProPublica nor Northpointe are charitably read as suggesting that the constitutive principles mentioning *separation* or *sufficiency* respectively exhaust *algorithmic fairness*. Both parties might well acknowledge further principles (Verma and Rubin 2018; Barocas, Hardt, and Narayanan 2019; Mehrabi et al. 2021) as constitutive. Even adding further principles of the sort just discussed might not suffice to capture the full range of phenomena related to *algorithmic fairness*. For example, Ovalle et al. (2023) show that *intersectionality* is not sufficiently understood in terms of standard categories of *algorithmic fairness*, which tend to elide social and/or historical conditions and power structures. The inconsistency thesis does not presuppose oversimplified, crude universalism because it is in principle compatible with constitutive principles bringing certain contextual specifics into focus. For example, one might be tempted to accept (e.g., in light of Ovalle et al. 2023, but also sources like Binns 2018, sect. 3.3) that

(2g) *algorithmic fairness* disapplies if in the assignment of risk scores of particular kinds, we fail to adjust for certain historical and/or structural injustices.

Rather than relying on overly truncated analyses of *algorithmic fairness*, the inconsistency thesis only requires that the principles leading to inconsistency are partly constitutive of the concept, that is, part of the set of constitutive principles.

One might think that the inconsistency thesis is old news. After all, the mentioned impossibility theorems have already established it. On reflection, however, while the theorems provide evidence to motivate the inconsistency thesis, they are not sufficient for it for at least two reasons. First, the theorems are typically portrayed as establishing that *algorithmic fairness* is "unachievable" (Berk et al. 2018) in practice. Unachievable scenarios can still be consistent. Second, the theorems themselves are silent on whether the principles shown to be irreconcilable are actually constitutive of *algorithmic fairness* (see 5.3). Overall, the inconsistency thesis implies that the news conveyed by the impossibility theorems are worse than many realize: the diagnosis of unachievability does not result from practical constraints or contingencies, but from inconsistencies inherent to the concept of *algorithmic fairness* itself.

Besides impossibility, it is sometimes argued that *algorithmic fairness* is inherently unstable. For example, Selbst et al. argue that the notion is "contextual", "contestable", and thus constantly "shifting" (Selbst et al. 2019, 61; see also: Hellman 2020; Birhane 2021). Whittlestone et al. highlight that for high-level AI ethics principles such as fairness, there are "tensions" not only between different principles, but also "conflicting meanings or values within a single principle" (2019, 198): relevant terms will be interpreted differently by different stakeholders, leading to mutually incompatible calls for action. The unifying feature of portrayals like these is that they claim ambiguity about the proper meaning of 'algorithmic fairness'—an ambiguity which requires deliberation in order to disambiguate or "arbitrate" (Selbst et al. 2019, 62) amongst different candidates. While potentially compatible with the inconsistency thesis, these views neither suggest nor anticipate that a plausible disambiguation might be inconsistent.

The inconsistency thesis has further implications for the status of disagreements and lingering contestation around *algorithmic fairness*. On the views just characterized, it is plausible to assume that under amenable discursive conditions, contestation and disagreement, for example around COMPAS, are virtuous in one sense because they indicate that discussants are engaged in an indispensable process of jointly specifying what fairness involves and requires. On the other hand, these dynamics also indicate that that process is still incomplete as discussants struggle to get clear about the concept. On the inconsistency thesis, the status of this struggle is not necessarily indicative of a shortcoming in grasping *algorithmic fairness*. On the contrary, these dynamics, even if irresolvable, circular, or inconclusive about whether *algorithmic fairness* applies in a given case, can in principle reflect understanding of *algorithmic fairness*, in particular that it can apply and disapply at the same time. Though an exploration of the ensuing practical implications is beyond the scope of the present paper, it is an important difference whether we are struggling to figure out fairness, or whether are we tracking what fairness is. In the former case, we need to try harder. In the latter case, we can proceed to potential downstream endeavors, such as fixing or replacing the inconsistent concept (as Scharp suggests), or figuring out how to shape our practices and technologies in view of inconsistency.

5. ALTERNATIVES TO THE INCONSISTENCY THESIS

As I have just argued, inconsistency arises for each of the foregoing pairs of constitutive principles of *algorithmic fairness*. I now turn to objections against the inconsistency thesis.

5.1 Joint Instantiation

A number of authors have recently put forward different strategies for tackling various forms of incompatibility amongst criteria of *algorithmic fairness*, in particular between *separation* and *sufficiency* (and/or relaxations thereof). As one example, Defrance and De Bie seek to complement the "negative result" of the impossibility theorem "with a positive one: a characterization of which combinations of fairness notions *are* possible" (2023, 851). They systematically investigate different combinations of seven different fairness criteria and show that besides the unattainable combinations figuring in the impossibility theorems, "12 maximal sets of these fairness measures are possible, among which are seven combinations of two measures, and five combinations of three measures" (2023, 851). A second strategy seeks to reconcile the criteria by way of reformulation: Beigang (2023) shows that if *separation* and *sufficiency* are rephrased and modified slightly by means of the method of "matching" familiar from drawing causal inferences from observational data, then at least under perfect matching conditions, the reformulated criteria are not only compatible, but even identical (Beigang 2023, 184–85). A third strategy operates with relaxation: Bell et al. show that if, instead of requiring equalization of fairness metrics across groups, we tolerate small margins (e.g., a 2% difference in the false positive rate for two groups), "it becomes possible to identify abundant sets of models that satisfy seemingly incompatible fairness constraints" (2023, 400). The putative obstacle to fairness raised by the theorems may thus be "overstated or even self-imposed" (2023, 409) and "achieving fairness along multiple metrics for multiple groups (and their intersections) is much more possible than was previously believed" (Bell et al. 2023, 400). The focus should shift towards exploring the space of solutions that jointly satisfy the relaxed constraints.

If successful, these proposals might give the initial appearance of undermining the inconsistency thesis. After all, each of them provides a workaround for escaping the prospect that application of *algorithmic fairness* on the basis of one constitutive principle triggers its disapplication on the basis of another

constitutive principle. In particular, satisfying *sufficiency* (as in 2a) would not necessarily entail that *separation* is undercut (as in 2b), and vice versa. This initial appearance, however, is mistaken. The inconsistency thesis can in principle be endorsed alongside these promising endeavors. Recall from illustrative inconsistent concepts such as *rable* or *truth* that inconsistent concepts need not lead into contradiction at each and every instance of deployment. *Truth* is an inconsistent concept (according to Scharp) which works perfectly well in ordinary circumstances. All that is needed for inconsistency is that there is a (class of) problem case(s), such as encounters with red tables or liar sentences, in which the concept leads into contradiction. Even if many combinations of fairness criteria are possible, this does not evade the inconsistency theses as long as there is one or more cases in which *algorithmic fairness* applies and disapplies at the same time. Even if joint instantiation of fairness measures initially taken to be incompatible is much more attainable than we thought, there will be circumstances in which *algorithmic fairness* seems to require precisely those combinations figuring in the impossibility theorems, and not (only) those mapped by Defrance and Debie. Likewise, there will be cases in which matching conditions relied on by Beigang's reconciliatory proposal are unavailable. And there will be cases in which the margins exploited by Bell et al. are intolerable. The existence of such cases vindicates the inconsistency thesis.

5.2 More Than One Concept

So far, the discussion has considered the thesis that 'the' concept of *algorithmic fairness* is inconsistent. However, that starting point might have been mistaken. David Chalmers (2011; 2020) defends conceptual pluralism: "there are multiple interesting concepts (corresponding to multiple interesting roles) in the vicinity of philosophical terms" (Chalmers 2011, 539). With the roles of a concept, Chalmers means the purposes it serves in our thought and practices. He criticizes that all too often, philosophers have debated the meaning of one philosophical term, while failing to realize that there is more than one concept in the vicinity. As a result, these philosophers have been talking past each other. Some of their disagreements dissolve once we disambiguate the specific concepts under consideration. For example, philosophers have been debating 'freedom' by using one and the same word, but without realizing that there is more than one concept in the vicinity: $freedom_1$ as the ability to do otherwise, $freedom_2$ as the ability to originate one's

own action, and so on. For Chalmers, "not much of substance depends on which [disambiguated concept] goes with the term" (Chalmers 2011, 539). More interestingly, some apparent disputes might dissolve once the respective claim under consideration is expressed by means of one or more of the disambiguated concepts (i.e., all disputants will then either agree or deny with the disambiguated thesis). For example, everyone might agree that *freedom*$_2$ is not required for moral responsibility. If so, this would be an insight into the roles and functions which this particular disambiguated concept does and does not play. If a dispute dissolves in such ways, the dispute was merely verbal rather than substantive.

Applied to *algorithmic fairness*, this framework appears to have at least the following implications for the inconsistency thesis. First, insofar as discussants fail to realize that there is a multiplicity of concepts in the vicinity of 'algorithmic fairness', they run the risk of talking past each other and engaging in merely verbal disputes as they are in fact concerned with slightly different concepts. Fortunately, various principles and criteria have already been distinguished in the debate. In view of this pluralism, some even criticize that misguided investments of research efforts into "bias busting" has led to a "proliferation" (Whittaker et al. 2018) of fairness formalizations. By now, the landscape is "littered with a multitude of measures" (Hellman 2020, 811). It is not obvious which elements of this multiplicity should be guiding, raising the need for "a pathway through that morass" (Hellman 2020, 815). In line with these diagnoses, Verma and Rubin (2018) collate a total of 20 different fairness criteria put forward in the algorithmic fairness discourse. Barocas et al. (2019, 75) present a list of 19 criteria, partially overlapping with Verma and Rubin's collection. Neither of them claims exhaustiveness. Mehrabi et al. (2021) distinguish 10 different definitions of fairness that are presupposed by discussants, if only implicitly. Hedden (2021) maps three different criteria for continuous risk scores, and eight criteria for binary predictions. The list could be continued. The emphases on contestation, and disagreement mentioned earlier (4.) also fit Chalmers's picture, as they can be construed as capturing the fact that discussants call for the satisfaction of different criteria respectively. Indeed, the COMPAS controversy is already one example in which the debate quickly focused on specific formalizations. All this suggests that just as Chalmers would expect, many concepts rather than just one are in play.

Second, just like Chalmers hopes, discussants have already begun to distinguish roles and functions of some of those specific concepts. For example, in view of the impossibility theorems, several authors attempt to disambiguate the scope of s*eparation* and *sufficiency* respectively. Hellman argues that *sufficiency* (i.e. conditionalizing on *R*) makes sense as a guide for "what one ought to believe about a scored individual" (2020, 812), whereas s*eparation* (i.e., conditionalizing on *Y*) relates more directly not (only) to what one ought to believe, but how one ought to act (2020, 835–39). This is intended to deflate Northpointe's emphasis on *sufficiency*, because COMPAS and many other cases of algorithmic (un)fairness relate to action, not (just) to belief. Further moves of disambiguating questions or aspects of fairness are made more or less explicitly in various framings. Berk et al. write that the impossibility theorems concern different "kinds of fairness" (Berk et al. 2018). Eva proposes that "it is crucial to keep track of distinctions between different kinds of unfairness, since the mechanisms that are best employed to combat or compensate for one kind of unfairness (e.g., the unjust historical origins of the correlations exploited by an algorithm) may not be effective in dealing with another kind of unfairness (e.g., an unfair statistical imbalance in the predictive tendencies of an algorithm)" (Eva 2022, 256). In other words, not only have various concepts been disambiguated; their particular roles, for example relative to other concepts or specific decision problems, receive attention as well.

Both aspects taken together put pressure on the inconsistency thesis. They suggest that the thesis concerns an unsuitable level of analysis: it is the result of aggregating a number of related but distinct concepts in the vicinity under one, putatively inconsistent concept. Chalmers thinks that inconsistency of philosophically relevant concepts is too rare to be of concern (Chalmers 2020, 15).

In view of these objections, proponents of the inconstancy thesis could investigate at least two broader questions. First, assuming conceptual pluralism, does it undermine the inconsistency thesis? One reason for caution is that even if a plurality of individually consistent concepts is distinguished, these might nevertheless figure in composite concepts that are inconsistent. For example, *separation* and *sufficiency* might figure in a composite concept (partly) constituted by (2a) and (2b). Rather than preventing inconsistency, pluralism actually leads to an explicit reiteration of a difficulty for evaluating whether or not

algorithmic fairness is consistent or inconsistent: the issue of specifying the constitutive principles (and the concepts figuring therein) that are to be assessed for consistency or inconsistency. If anything, the high number of criteria identified by pluralism broadens the range of prima facie eligible constitutive principles, and thus could invite rather than mitigate joint inconsistency.

Second, how plausible is pluralism about algorithmic fairness? The "littering" charge according to which there are now too many criteria to do sensible work can be deflected at least to some extent. As Barocas et al. (2019, 45–75) show, each of the 19 criteria they identify is in fact either equivalent to or a relaxation of three fundamental criteria: s*eparation*, s*ufficiency*, and *independence* (which requires that random variables (A,R) satisfy $A \perp R$). Admittedly, this is not perfect unity, but it is prima facie evidence against unconstrained pluralism and littering charges. Van Nood and Yeomans even argue that all fairness definitions are operationalizations of one and the same, unifying claim: that like cases ought to be treated alike. The various fairness definitions are "tests for ensuring that groups that ought to be treated as alike are in fact so treated" (van Nood and Yeomans 2021). Likewise, Holm posits that different fairness criteria "can be understood as applications of a single principle of fairness […] according to which fairness in the distribution of a good between people consists in the proportional satisfaction of their claims to the good" (2023). Though not identical in substance, a similarity between these positions is that *pace* pluralism, each of them considers the plurality of fairness operationalizations as rooted in one, single idea. And as the inconsistency thesis highlights, in each of these cases we might scrutinize that single idea for consistency or inconsistency.

5.3 Not Constitutive

A further, serious challenge to the inconsistency thesis is that the principles that trigger simultaneous application and disapplication of *algorithmic fairness* might not be even partly constitutive of the concept.

5.3.1 Counterexamples to the principles

Various counterexamples to different fairness criteria have been proposed. For example, Hedden provides a thought experiment involving a seemingly innocuous algorithm for classifying people in different rooms based on the prediction of coin tosses (Hedden 2021). He observes that the envisioned algorithm undercuts

10 different algorithmic fairness criteria, including *separation*. The key point for him is: "[i]t should be clear that these facts do not show that the predictive algorithm was unfair or biased against any individuals in virtue of their being members of one room or the other. The algorithm is perfectly fair; it is the statistical criteria that must go" (Hedden 2021, 222). He would thus reject (2b).

Similarly, the case of redlining (3.2), which was presented as bringing out an inconsistency between (2a) and (2c), can be read differently: rather than suggesting an inconsistency between the two proposed principles, the obvious unfairness of redlining can be seen as a reductio of (2a). *Sufficiency* is satisfied, yet a redlining algorithm is unfair. (2a) thus cannot be constitutive of algorithmic fairness (Corbett-Davies et al. 2017, 803–4; Eva 2022, 253–55).

The counterexamples are also contested (e.g., Viganò et al. 2022). This is not the place to evaluate these debates, only to note that the inconsistency thesis is not tied to a particular pair of constitutive principles that trigger simultaneous application and disapplication of *algorithmic fairness*. All the thesis requires is that one such pair is constitutive of the concept.

5.3.2 Inconsistent beliefs

Another challenge concerns the locus of inconsistency. According to Herman Cappelen, who is critical of Scharp's framework (2.), concepts of philosophical or practical interest are almost never inconsistent. Instead, the appearance of inconsistency stems solely from inconsistent beliefs about, or conceptions of, the respective concept (Cappelen 2018, 86–88). Their inconsistency shows nothing about the nature of the concept itself, which might well be perfectly consistent after all.

According to Cappelen, the primary source of the inconsistency of relevant beliefs is "metasemantic messiness", that is, complexity in how the meaning of expressions is determined and changes over time (Cappelen 2018, 57, 88–91), "that leads to a term's having defective semantic content and to defective thoughts" (Cappelen 2018, 91). In this process, thinkers run the risk of being misled into at least some false beliefs about the subject matter. Amongst those beliefs are putative constitutive principles of seemingly inconsistent concepts—principles which on reflection, for example in view of the counterexamples just mentioned, should not actually be seen as constitutive of the respective concept.

Indeed, *algorithmic fairness* is a complex concept with a rich history (Hutchinson and Mitchell 2019). As mentioned above (4.), it continues to generate debate and disagreement. It would be surprising if these processes had not led to at least some level of metasemantic messiness, some of which might have led to misguided thoughts that brought about the appearance of inconsistency. Cappelen's critical alternative view thus challenges anyone who considers the inconsistency thesis to determine whether any apparent inconsistency does not just reside at the level of our beliefs, but genuinely reflects the concept itself.

This general challenge does not uniquely apply to the inconsistency thesis about *algorithmic fairness*. Together with the counterexamples, it suggests that pending a resolution of these issues, the inconsistency thesis should be restricted to a conditional claim: *if* one or more of the pairs of principles discussed above (3.) are seen as (partly) constitutive, then they lead into contradiction and suggest that *algorithmic fairness* is an inconsistent concept.

6. CONCLUSION

Inconsistent concepts are concepts that can apply and disapply at the same time, leading to defective thoughts (2.). A range of findings on the relation between different criteria of *algorithmic fairness* suggest that the concept is inconsistent (3.). While the inconsistency thesis is in principle compatible with a range of extant understandings of *algorithmic fairness*, it denotes a distinctive option in argumentative space (4.). Amongst others, it adds an important clarification on why we observe tensions and contestations around the concept: not because we arbitrate between different, mutually incompatible fairness notions, but because we (attempt to) track one and the same inconsistent concept. While there are various ways to resist this inconsistency thesis (5.), there is good reason to endorse the conditional claim that if popular criteria of algorithmic fairness are seen as constitutive of the concept, then this concept is inconsistent.

As should have become apparent, I have not argued that it is tangibly useful or practically beneficial to think of *algorithmic fairness* as inconsistent. Instead, I have suggested that we have to reckon with the possibility that the inconsistency thesis is true. These findings raise various normative questions for further research. Given the societal impact of algorithms on the one hand, and the prominent role of principle-based guidance in AI ethics on the other, there is a need to investigate whether and how the field could detect, rule

out, recognize, and/or account for inherent inconsistency (as opposed to, e.g., mere ambiguity) in a key concept figuring in a large number of its foundational and implementation-oriented works. Once inconsistency is ascertained, the question arises what we should do about it. Should we fix or replace the concept? If so, how? Meanwhile, the inconsistency thesis does not entail that we lack reasons to attempt instantiations of fairness, but might motivate going about such endeavors in particular ways. Throughout, it will remain a pressing question which methods might help identifying and mitigating any defects in our thoughts about the subject matter that the concept has led us into by virtue of its inconsistency. For all these endeavors, a necessary condition is an articulation and tentative evaluation of the inconsistency thesis—a project on which the present paper has put forward a suggestion.

Patrik Hummel

Philosophy & Ethics Group

Department of Industrial Engineering & Innovation Sciences

Eindhoven University of Technology

De Zaale, Eindhoven, The Netherlands

p.a.hummel@tue.nl

ACKNOWLEDGEMENTS

This work is part of the research program Ethics of Socially Disruptive Technologies (ESDiT), which is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organisation for Scientific Research (NWO grant number 024.004.031). For feedback on earlier versions of this text, I am grateful to Kevin Scharp, the ESDiT "The Future of a Fair & Free Society" research line, anonymous referees for ACM FAccT and the American Philosophical Quarterly, participants at the "Political Hope Workshop vol. 1" at the Chair of Social Ethics, University of Bonn, and my colleagues in the TU/e Philosophy & Ethics Group.

REFERENCES

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. “Machine Bias.” ProPublica. 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Beigang, Fabian. 2023. “Reconciling Algorithmic Fairness Criteria.” *Philosophy & Public Affairs* 51 (2): 166–90. https://doi.org/10.1111/papa.12233.

Bell, Andrew, Lucius Bynum, Nazarii Drushchak, Tetiana Zakharchenko, Lucas Rosenblatt, and Julia Stoyanovich. 2023. “The Possibility of Fairness: Revisiting the Impossibility Theorem in Practice.” In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 400–422. FAccT ’23. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3593013.3594007.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. “Fairness in Criminal Justice Risk Assessments: The State of the Art.” *Sociological Methods & Research* 50 (1): 3–44. https://doi.org/10.1177/0049124118782533.

Binns, Reuben. 2018. “Fairness in Machine Learning: Lessons from Political Philosophy.” In *Conference on Fairness, Accountability and Transparency*, 149–59. PMLR. http://proceedings.mlr.press/v81/binns18a.html.

Birhane, Abeba. 2021. “Algorithmic Injustice: A Relational Ethics Approach.” *Patterns* 2 (2). https://doi.org/10.1016/j.patter.2021.100205.

Burrell, Jenna, and Marion Fourcade. 2021. “The Society of Algorithms.” *Annual Review of Sociology* 47 (1): 213–37. https://doi.org/10.1146/annurev-soc-090820-020800.

Cappelen, Herman. 2018. *Fixing Language: An Essay on Conceptual Engineering*. First edition. Oxford, United Kingdom ; New York, NY: Oxford University Press.

Chalmers, David J. 2011. “Verbal Disputes.” *The Philosophical Review* 120 (4): 515–66. https://doi.org/10.1215/00318108-1334478.

———. 2020. “What Is Conceptual Engineering and What Should It Be?” *Inquiry*, September, 1–18. https://doi.org/10.1080/0020174X.2020.1817141.

Chouldechova, Alexandra. 2017. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.” *Big Data* 5 (2): 153–63. https://doi.org/10.1089/big.2016.0047.

Corbett-Davies, Sam, and Sharad Goel. 2018. “The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning.” *arXiv:1808.00023 [Cs]*, August. http://arxiv.org/abs/1808.00023.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. “Algorithmic Decision Making and the Cost of Fairness.” In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797–806. Halifax NS Canada: ACM. https://doi.org/10.1145/3097983.3098095.

Defrance, Marybeth, and Tijl De Bie. 2023. “Maximal Fairness.” In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 851–80. FAccT ’23. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3593013.3594048.

Dieterich, William, Christina Mendoza, and Tim Brennan. 2016. *COMPAS Risk Scales: Demonstratig Accuracy Equity and Predictive Parity*. Northpointe Inc. Research Department.

Eva, Benjamin. 2022. “Algorithmic Fairness and Base Rate Tracking.” *Philosophy & Public Affairs* 50 (2): 239–66. https://doi.org/10.1111/papa.12211.

Fleisher, Will. 2021. “Evidence of Fairness: On the Uses and Limitations of Statistical Fairness Criteria.” *SSRN*. https://doi.org/10.2139/ssrn.3974963.

Haslanger, Sally. 2000. “Gender and Race: (What) Are They? (What) Do We Want Them to Be?” *Noûs* 34 (1): 31–55.

———. 2012. *Resisting Reality: Social Construction and Social Critique*. New York: Oxford University Press.

Hedden, Brian. 2021. “On Statistical Criteria of Algorithmic Fairness.” *Philosophy & Public Affairs* 49 (2): 209–31. https://doi.org/10.1111/papa.12189.
Hellman, Deborah. 2020. “Measuring Algorithmic Fairness.” *Virginia Law Review* 106 (4): 811–66.
Holm, Sune. 2023. “Egalitarianism and Algorithmic Fairness.” *Philosophy & Technology* 36 (1): 6. https://doi.org/10.1007/s13347-023-00607-w.
Hopster, Jeroen. 2021. “What Are Socially Disruptive Technologies?” *Technology in Society* 67 (November): 101750. https://doi.org/10.1016/j.techsoc.2021.101750.
Hughes, George Edward, and Max Cresswell. 1996. *A New Introduction to Modal Logic*. London ; New York: Routledge.
Hutchinson, Ben, and Margaret Mitchell. 2019. “50 Years of Test (Un)Fairness: Lessons for Machine Learning.” In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 49–58. FAT* ’19. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287600.
Kearns, Michael, and Aaron Roth. 2020. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. New York: Oxford University Press.
Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2017. “Inherent Trade-Offs in the Fair Determination of Risk Scores.” *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. https://doi.org/10.4230/LIPICS.ITCS.2017.43.
Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. “How We Analyzed the COMPAS Recidivism Algorithm.” ProPublica. 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.
Löhr, Guido. 2023. “Conceptual Disruption and 21st Century Technologies: A Framework.” *Technology in Society* 74 (August): 102327. https://doi.org/10.1016/j.techsoc.2023.102327.
Margolis, Eric, and Stephen Laurence. 2022. “Concepts.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Fall 2022. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2022/entries/concepts/.
Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. “A Survey on Bias and Fairness in Machine Learning.” *ACM Computing Surveys* 54 (6): 1–35. https://doi.org/10.1145/3457607.
Mitchell, Shira, Eric Potash, Solon Barocas, Alexander D’Amour, and Kristian Lum. 2021. “Algorithmic Fairness: Choices, Assumptions, and Definitions.” *Annual Review of Statistics and Its Application* 8 (1): 141–63. https://doi.org/10.1146/annurev-statistics-042720-125902.
Nood, Ryan van, and Christopher Yeomans. 2021. “Fairness as Equal Concession: Critical Remarks on Fair AI.” *Science and Engineering Ethics* 27 (6): 73. https://doi.org/10.1007/s11948-021-00348-z.
Northpointe. 2015. *Practitioner’s Guide to COMPAS Core*.
Ovalle, Anaelia, Arjun Subramonian, Vagrant Gautam, Gilbert Gee, and Kai-Wei Chang. 2023. “Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness.” In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 496–511. AIES ’23. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3600211.3604705.
Priest, Graham, Francesco Berto, and Zach Weber. 2022. “Dialetheism.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Fall 2022. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2022/entries/dialetheism/.
Saxena, Nripsuta Ani, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. 2019. “How Do Fairness Definitions Fare?: Examining Public Attitudes Towards Algorithmic Definitions of Fairness.” In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 99–106. Honolulu HI USA: ACM. https://doi.org/10.1145/3306618.3314248.
Scharp, Kevin. 2007. “Replacing Truth.” *Inquiry* 50 (6): 606–21. https://doi.org/10.1080/00201740701698589.
———. 2013. *Replacing Truth*. Oxford: Oxford University Press.

———. 2020. “Philosophy as the Study of Defective Concepts.” In *Conceptual Engineering and Conceptual Ethics*, edited by Herman Cappelen, David Plunkett, and Alexis Burgess, 396–416. Oxford University Press. https://doi.org/10.1093/oso/9780198801856.003.0001.

Selbst, Andrew D., Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. “Fairness and Abstraction in Sociotechnical Systems.” In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. Atlanta GA USA: ACM. https://doi.org/10.1145/3287560.3287598.

Sider, Theodore. 2010. *Logic for Philosophy*. Oxford ; New York: Oxford University Press.

Tsamados, Andreas, Nikita Aggarwal, Josh Cowls, Jessica Morley, Huw Roberts, Mariarosaria Taddeo, and Luciano Floridi. 2022. “The Ethics of Algorithms: Key Problems and Solutions.” *AI & SOCIETY* 37 (1): 215–30. https://doi.org/10.1007/s00146-021-01154-8.

Verma, Sahil, and Julia Rubin. 2018. “Fairness Definitions Explained.” In *Proceedings of the International Workshop on Software Fairness*, 1–7. Gothenburg Sweden: ACM. https://doi.org/10.1145/3194770.3194776.

Viganò, Eleonora, Corinna Hertweck, Christoph Heitz, and Michele Loi. 2022. “People Are Not Coins: Morally Distinct Types of Predictions Necessitate Different Fairness Constraints.” In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2293–2301. Seoul Republic of Korea: ACM. https://doi.org/10.1145/3531146.3534643.

Whittaker, Meredith, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Mysers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. 2018. *AI Now Report 2018*. New York: AI Now Institute at New York University.

Whittlestone, Jess, Rune Nyrup, Anna Alexandrova, and Stephen Cave. 2019. “The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions.” In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 195–200. Honolulu HI USA: ACM. https://doi.org/10.1145/3306618.3314289.