DOI:10.1145/3433949

What does it mean to be fair?

BY SORELLE A. FRIEDLER, CARLOS SCHEIDEGGER, AND SURESH VENKATASUBRAMANIAN

The (Im)possibility of Fairness:

Different Value Systems Require Different Mechanisms For Fair Decision Making

AUTOMATED DECISION-MAKING SYSTEMS (often machine learning-based) now commonly determine criminal sentences, hiring choices, and loan applications. This widespread deployment is concerning, since these systems have the potential to discriminate against people based on their demographic characteristics. Current sentencing risk assessments are racially biased,[4] and job advertisements discriminate on gender.[8] These concerns have led to an explosive growth in fairness-aware machine learning, a field that aims to enable algorithmic systems that are fair by design.

To design fair systems, we must agree precisely on what it means to be fair. One such definition is *individual fairness*:[10] individuals who are similar (with respect to some task) should be treated similarly (with respect to that task). Simultaneously, a different definition states that demographic groups should, on the whole, receive similar decisions. This *group fairness* definition is inspired by civil rights law in the U.S.[5,11] and U.K.[21] Other definitions state that fair systems should err evenly across demographic groups.[7,13,24] Many of these definitions have been incorporated into machine learning pipelines.[1,6,11,16,25]

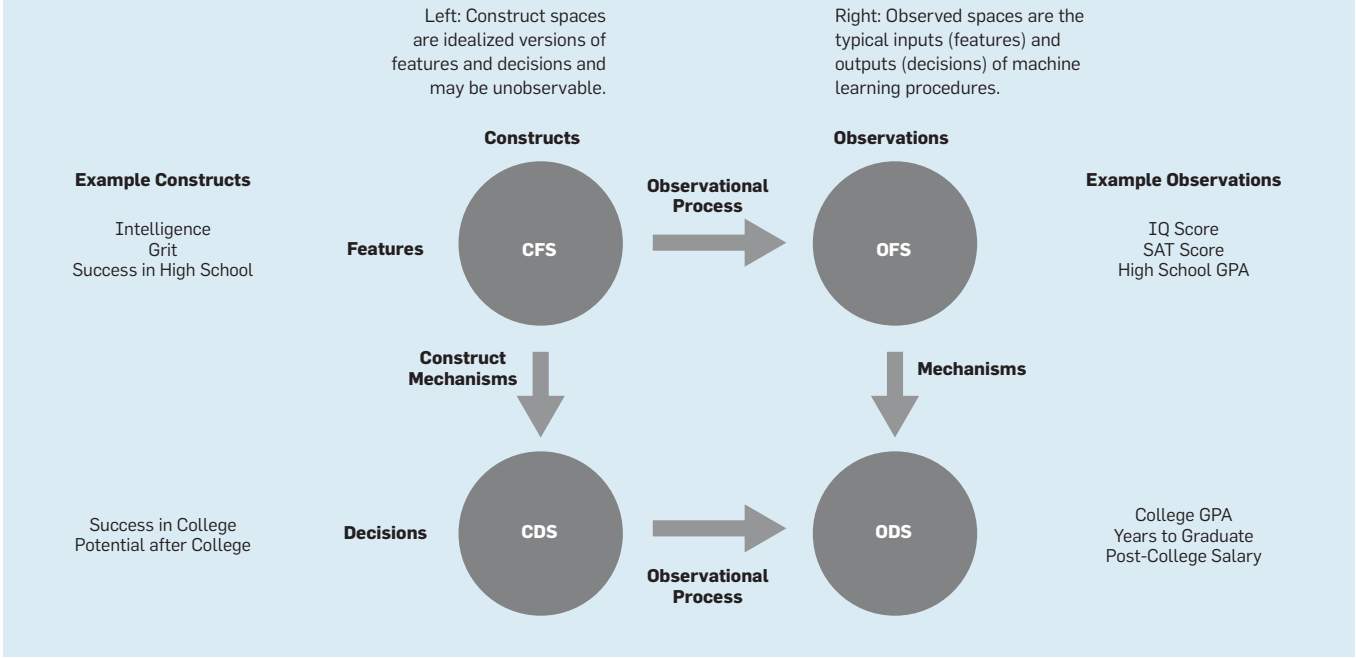
In this article, we introduce a framework for understanding these different definitions of fairness and how they relate to each other. Crucially, our framework shows these definitions and their implementations correspond to different axiomatic beliefs about the world. We present two such *worldviews* and will show they are fundamentally incompatible. First, one can believe the observation processes that generate data for machine learning are structurally biased. This belief provides a justification for seeking non-discrimination. When one believes that demographic groups are, on the whole, fundamentally similar, group fairness mechanisms successfully guarantee the top-level goal of non-discrimination: similar groups receiving similar treatment. Alternatively, one can assume the observed data generally re-

» key insights

- **The world is structurally biased and makes structurally biased data. Observation is a process. When we create data, we choose what to look for.**
- **Every automated system encodes a value judgment. Accepting training data as given implies structural bias does not appear in the data and that replicating the data as given would be just.**
- **Different value judgments can require satisfying contradicting fairness properties each leading to different societal outcomes.**
- **Researchers and practitioners must document data collection processes, worldviews, and value assumptions.**
- **Value decisions must come from domain experts and affected populations; data scientists should listen to them in order to build in values that lead to justice.**

IMAGE BY ANDRIJ BORYS ASSOCIATES, USING SHUTTERSTOCK

Figure 1. Our model of algorithmic decision making involves transformations among four spaces.



flects the true underlying reality about differences between people. These worldviews are in conflict; a single algorithm cannot satisfy either definition of fairness under both worldviews. Thus, researchers and practitioners ought to be intentional and explicit about worldviews and value assumptions: the systems they design will always encode some belief about the world.

An Example

We illustrate the practice of fairness in decision making with the example of a college admissions process. We often think of this process as starting with the input provided by application materials and ending with an admittance decision. Here, we will take a broader view of the process, including the goals of the admissions office and an assessment of the resulting decisions. In this broader view, the first step of the process is determining a set of idealized features to be used in an admissions decision.

For example, some consider personal qualities such as self-control, growth mindset, and grit to be determining factors in later success.[3] *Grit* is roughly defined as an individual's ability to demonstrate passion, perseverance, and resilience toward their chosen goal. An admissions committee could decide to use grit as an idealized predictor. However, grit (and other idealized features) are not directly observable; they are measured indirectly and imprecisely through self-reported surveys and other proxies.[9]

Similarly, when considering admissions decisions, it is also important to determine what an idealized decision should be predicting. An admissions office might decide that decisions should be made based on the predicted potential of an applicant. Since "potential" is unobservable, systems might use more directly measurable—and more problematic—features such as college GPA upon graduation.

This college admissions example demonstrates the basic pipeline in human decision-making systems, which is also mimicked in algorithmic systems. We decide on idealized features, and measure observed, possibly flawed versions of these features. We determine an idealized prediction goal, and measure observed features to predict an observed goal. We next formalize this decision-making framework.

Spaces: Construct vs. Observed and Features vs. Decisions

We model an algorithm making decisions about individuals as a mapping from a space of information about people, which we will call a *feature space*, to a space of decisions, which we will call a *decision space*. We assume each space is a collection of information about people endowed with a *distance metric* (specifically, a function of pairs of elements that satisfies reflexivity, symmetry, and the triangle inequality). This reflects that a process for discovering mappings from features to decisions usually exploits the *geometry* of these spaces.

We introduce two types of spaces: *construct spaces* and *observed spaces*. Construct spaces contain an idealized representation of information about people and decisions. These spaces may include unmeasurable "constructs" (for example, grit). Observed spaces contain the results of an observational process that maps information about people or decisions to measurable spaces of inputs or outputs (for example, the results of self-reported surveys designed to measure grit). An observational process can be noisy, including additional information not found in the associated construct space, missing information, or even containing a large distance skew in the mapping. Observational processes don't have to maintain information that is useful for the decision-making task.

These two distinctions—between feature and decision spaces and between construct and observed spaces—naturally give rise to four spaces that we claim are necessary for analyzing the fairness of a decision-making procedure (as illustrated in Figure 1):

The Construct Feature Space (*CFS*) is the space representing the "desired"

or "true" collection of information about people to use as input to a decision-making procedure. For example, this includes features like intelligence or grit for college admission.

The Observed Feature Space (*OFS*) is the space containing the observed information about people, generated by an observational process $g : CFS \rightarrow OFS$ that generates an entity $\hat{p} = g(p)$ from a person $p \in CFS$. For example, this includes the results of standardized tests or personal essays.

The Construct Decision Space (*CDS*) is the space representing the idealized outcomes of a decision-making procedure. For example, this includes how well a student will do in college.

The Observed Decision Space (*ODS*) is the space containing the per-person observed decisions from a concrete decision-making procedure, generated by an observational process mapping $CDS \rightarrow ODS$. For example, this includes the GPA of a student after their freshman year.

To understand the interactions between these spaces, we start with a prediction task, determine an idealized decision goal, posit features that seem to control the decision, and then imagine ways of measuring those features and decisions. Explicitly considering the existence of the construct space is rare in practice; we argue that explicit goals and assumptions are necessary when considering fairness. It is worth emphasizing here that the construct spaces represent our best *current* understanding of the underlying factors involved in a task rather than some kind of Platonic universal ideal. They are therefore *contingent* on the current specific ideas and best practices about how to make the decision in the given context.

TL;DR

Constructs are the idealized features and decisions we wish we could use for decision-making.

Observed features and decisions are the measurable features and outcomes that are actually used to make decisions.

These **may be different**, and it is important to be explicit about the distinction.

Fairness and Non-Discrimination

Traditional data science and machine learning can be understood as focusing on creating transformations between the observed feature and observed decision spaces. These mechanisms are used in real-world decision-making practices by taking observed data as input. On the other hand, fairness is defined as a property of an idealized *construct mechanism* that maps individuals to construct decisions based on their construct features. The goal of algorithmic fairness is to develop real-world mechanisms that match the decisions generated by these construct mechanisms. In order to discuss these fairness-aware mechanisms further, we first describe different notions of fairness within our framework.

Individual fairness. Since fairness is an idealized property operating based on underlying and potentially unobservable information about people, it is most natural to define it within the construct spaces. The definition of fairness is task specific and prescribes desirable (potentially unobservable) outcomes for a task in the construct decision space. Since the solution to a task is a mapping from the construct feature space to the construct decision space, a definition of fairness should describe the properties of such a construct mechanism. Inspired by the fairness definition due to Dwork et al.,[20] we define individual fairness as follows:

Individual fairness. Individuals who are similar (with respect to the task) in the *CFS* should receive similar decisions in the *CDS*.

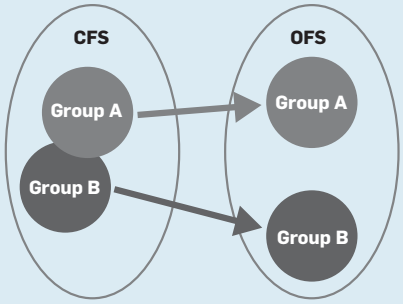
TL;DR

Individual fairness is the goal of giving similar individuals similar decisions.

Non-discrimination. While our fairness definition focuses on the individual, there are often groups that share characteristics (such as race, gender, and so on) and fairness should be considered with respect to these group characteristics (or combinations of them) as well. Group membership is often defined on innate, culturally defined, immutable characteristics, or those protected for historical reasons. It is often considered unacceptable (and sometimes illegal) to use group membership as part of a decision-making process. Thus, we define non-discrimination as follows:

Figure 2. An illustration of group skew in the mapping between the feature spaces.

The mapping moves the groups further apart in the observed space than they were in the construct space, increasing the inter-group distance while maintaining the intra-group distance.



Non-discrimination. Groups who are similar (with respect to the task) in the *CFS* should, as a whole, receive similar decisions in the *CDS*.

In this work, consider group membership as a characteristic of an individual; thus, each of the four spaces admits a partition into groups, induced by the group memberships of individuals represented in these spaces.

Discrimination manifests itself in unequal treatment of groups. To quantify this, we first introduce the notion of *group skew*. Given two spaces and their associated group partitioning, the group skew of the mapping between the spaces is the extent to which the groups are, as a whole, mapped differently from each other. An illustration of group skew between the feature spaces is given in Figure 2. The goal in any formalization of group skew is to capture the relative difference in the mappings of groups with respect to each other, rather than (for example) a scaling transformation that transforms all groups the same way. This can be thought of as quantifying any difference in treatment based on group membership.[a] Thus, nondiscrimination is defined as the lack of group skew in the mapping from *CFS* to *CDS*. This notion attempts to capture the idea of fair treatment of groups.

a We introduce one possible geometric formalization of group skew in Friedler et al.[12]

TL;DR

Non-discrimination is the goal of giving similar groups on the whole similar decisions.

While non-discrimination shares many characteristics with the notions of group fairness that previous work has studied, an important distinction that we introduce here is that non-discrimination is defined as a property of the mapping between the *construct* spaces, while group fairness is generally defined in practice via *group fairness mechanisms* that restrict mappings between the *observed* feature and decision spaces. The ideas we develop next will allow us to understand that mechanisms that guarantee group fairness notions are generally doing so with the goal of guaranteeing non-discrimination, but formalizing the relationship between these two notions requires axiomatic assumptions about the world.

Worldviews and Assumptions

Fairness goals are defined as properties of construct mechanisms. Real-world decision making, however, must use mechanisms that map between the observed spaces. Thus, fair algorithm designers are forced to make assumptions about the observational processes mapping from construct to observed spaces in order to make real-world decisions.

We now describe two such axiomatic assumptions that are motivated by associated worldviews. While we introduce these two axioms as different worldviews or belief systems, they can also be strategic choices. Roemer identifies the goal of such choices as ensuring that negative attributes due to an individual's circumstances of birth or to random chance should not be held against them, while individuals should be held accountable for their effort and choices.[20] In our framework, this translates to a decision of which axiom to choose. In our college admissions example, this may be the difference between asserting the admissions process should serve as a social equalizer, so that, for example, applicants from different class backgrounds are admitted at approximately the same rate, or believing that features such as GPA and SAT scores accurately reflect effort and understanding.

Worldview: What you see is what you get. One worldview handles the uncertainty of the observational process mapping between any construct and observed space by asserting that the spaces are *essentially the same* with respect to the task.

What You See Is What You Get (WYSIWYG): The observational process between a given construct space and observed space maintains the relative position of individuals with respect to the task.

It is common in data science to directly use any observed data that is available; doing so without modification or further evaluation is an adoption of the WYSIWYG worldview and assumption. Importantly, a consequence of understanding that WYSIWYG is an axiomatic assumption is that we can reevaluate this assumption.

Worldview: Structural bias. What if the construct space is not accurately represented by the observed space? In many real-world applications, the transformation from construct to observed space is non-uniform in a societally biased way. To capture this idea, we return to the notion of group skew from the previous section. We define structural bias as group skew between construct and observed spaces, that is, an observational process that treats groups differently. Such a process is illustrated in Figure 2.

In the cases where observed data is believed to suffer from structural bias, a common fairness goal is non-discrimination—a goal that aims to make sure that the group skew that is present in the observed data is not found in the resulting decisions. Unfortunately, non-discrimination is difficult to achieve directly since it is defined in terms of the construct spaces, and the existence of structural bias precludes us from using the observed spaces as a reasonable representation of the construct spaces (unlike the WYSIWYG worldview).

Instead, a common underlying assumption of this worldview is that in one or both of the construct spaces *all groups look essentially the same*. It asserts there are no innate differences between demographic groups. There may still be variation between individuals within the group, but the assumption is that on the whole, for example, as a distribution, the groups are essentially the same.

We're All Equal (WAE). Within a given construct space all groups are essentially the same.

This axiom could be applied to either or both of the construct spaces (*CFS* and *CDS*). It appears implicitly in much of the literature on statistical discrimination and disparate impact.

There is an alternate interpretation of this axiom: the groups are not necessarily equal, but for the purposes of the decision-making process, we would prefer to treat them as if they were. In this interpretation,[20] any difference in the groups' performance (for example, academic achievement) is due to factors outside their individual control (for example, the quality of their neighborhood school) and should not be taken into account in the decision-making process. This interpretation has the same mathematical outcome as if groups are assumed equal, and thus a single axiom covers both these interpretations.

We reiterate that structural bias—the way in which observations are systematically distorted—is separate from the WAE axiom, which is a device that allows us to interpret observed skew as a measure of structural bias.

TL;DR

Any attempt to design fair decision making is **forced to make assumptions** about the observational process and/or construct space. There are two main such assumptions:

Work that **uses the observed data directly** is making a WYSIWYG assumptions; and,

Work that attempts to guarantee statistical parity and other **group fairness notions as a measure** is making a WAE assumption.

Later, we will explore the ways that different works have made these assumptions further.

Consequences

We now sketch some consequences of attempts to achieve individual fairness and non-discrimination under different worldviews. A more detailed analysis can be found in our extended work;[12] other authors have also started to build on the framework we lay out.[23]

Mechanisms achieve the goals of their worldviews. How can individual fairness be achieved? *Individual fairness mechanisms* are algorithms that guarantee that individuals who are similar in the observed feature space receive similar decisions in the observed decision space. Under WYSIWYG assumptions on both the feature and decision observational processes, individual fairness mechanisms can be shown to guarantee individual fairness (via function composition).

Group fairness mechanisms ensure that groups are mapped to, on the whole, similar decisions in the observed decision space. Under a WAE assumption (applied to the *CFS*), group fairness mechanisms can be shown to guarantee non-discrimination since groups are assumed to be essentially the same in the construct feature space and the mechanism guarantees that this is enforced in the mapping to the observed decision space.

TL;DR

Under a **WYSIWYG** assumption, **individual fairness** can be guaranteed.

Under a **WAE** assumption, **non-discrimination** can be guaranteed.

Conflicting worldviews necessitate different mechanisms. Do mechanisms exist that can guarantee individual fairness or non-discrimination under both worldviews?

Unfortunately, WYSIWYG appears to be crucial to ensuring individual fairness: if there is structural bias in the decision pipeline, no mechanism can guarantee individual fairness. Fairness can only be achieved under the WYSIWYG worldview using an individual fairness mechanism and using a group fairness mechanism will be *unfair* within this worldview.

What about non-discrimination? Unfortunately, another counterexample shows these mechanisms are not agnostic to worldview. Suppose that the construct and observed decision spaces are the same and that two groups are very far apart in the *CFS* with images in the *OFS* that are even further apart. Applying an individual fairness mechanism to the *OFS* will result in decisions that preference the group that performed better with respect to the task in the *CFS* more than is warranted compared to the other group; this is discriminatory.

Researchers and practitioners ought to be intentional and explicit about worldviews and value assumptions—the systems they design will always encode some belief about the world.

Choice in mechanism must thus be tied to an explicit choice in worldview. Under a WYSIWYG worldview, only individual fairness mechanisms achieve fairness (and group fairness mechanisms are unfair). Under a structural bias worldview, only group fairness mechanisms achieve non-discrimination (and individual fairness mechanisms are discriminatory).

TL;DR

Fairness-aware algorithms cannot guarantee fairness or non-discrimination under *both* the WYSIWYG and structural bias worldviews. Choice in algorithms must be tied to an **explicit choice in worldview**.

Placing Literature in Context

Our framework lets us analyze existing literature in fairness (for broader surveys, see Romei et al.[21] and Zliobaite[27]) to see what axiomatic positions different solutions might implicitly be taking. For this analysis, we distinguish papers that propose new fairness measures and/or interventions from the smaller number of papers that provide a metanalysis of fairness definitions.

Fairness measures and algorithms. Our findings are twofold. First, we find we can categorize existing work based on fairness *measure* and associated assumption on the *decision spaces*. Measures that assume that existing decisions are correct and optimize fairness conditioned on that assumption adopt the WYSIWYG axiom between decision spaces. Measures that are open to changing the observed decisions in the data adopt the WAE axiom. Second, once we categorize measures based on the decision space axiomatic choice, we can categorize algorithms based on axioms governing the feature spaces. Algorithms that work to change the data representation adopt a viewpoint that the data may not be correct, and generally do this according to the WAE axiom. Algorithms that make no change to the data before optimizing for a measure implicitly make the WYSIWYG axiomatic assumption between feature spaces.

Note that it is not a contradiction to have an algorithm that, based on its measure, assumes WAE in the construct decision space and WYSIWYG

between the feature spaces. In fact, many algorithms make these dual assumptions in practice.

Early work on non-discrimination that initiated the study of fairness-aware data mining considered the difference in outcomes between groups. Specifically, let $Pr[C = \text{Yes} \mid G = 0]$ be the probability of people in the unprivileged group receiving a positive classification and $Pr[C = YES \mid G = 1]$ be the probability of people in the privileged group receiving a positive classification. Calders and Verwer[6] introduce the idea of a *discrimination score* defined as $Pr[C = YES \mid G = 1] - Pr[C = YES \mid G = 0]$. Their goal, and the goal of much subsequent work also focusing on this measure,[14,16,22,26] was to bring this difference to zero. The assumption here is that groups should, as a whole, receive similar outcomes. The implicit assumption is that the original decisions received by the groups (that is, the decisions used for training) may not be correct if this difference is not small. This reflects an underlying WAE axiom in the construct decision space. The four-fifths rule for disparate impact[5,11,25] focuses on a similar measure (taking the ratio instead of the difference) and also assumes the WAE axiom.

A 2016 ProPublica study[4] examined the predicted risk scores assigned to defendants by the COMPAS algorithm and found that Black defendants were about twice as likely to receive incorrect high-risk scores (bad errors), while White defendants were about twice as likely to receive incorrect low risk scores (good errors). This inspired the development of measures for equalizing the group conditioned error rates of algorithms (termed "equal odds"[13] or "disparate mistreatment"[24]), with the idea that different groups should receive the same impact of the algorithm conditioned on their outcomes. This implicitly assumes the observed outcomes (observed decision space) reflect true decisions, that is, these measures assume the WYSIWYG axiom between the decision spaces. It is interesting that while these classes of measures are all considered group fairness measures, they make different assumptions about the decision spaces.

Axiomatic assumptions about feature spaces are determined by the choice of algorithm. Some works attempt to ensure non-discrimination by modifying the decision algorithm[6,16] while others change the outcomes after the decision has been drafted.[15] Even though these algorithms try to ensure non-discrimination and assume the WAE axiom in the construct decision space, they implicitly assume the WYSIWYG axiom between the feature spaces by using training data without modification. Algorithms for achieving fairness on group-conditioned error rates[13,24] focus on constraining this measure using the observed data as given, so these algorithms assume WYSIWYG between the feature spaces. Given that these algorithms focused on the group-conditioned error rate also assume the WYSIWYG between decision spaces, we claim that they adopt the WYSIWYG worldview and not a structural bias worldview.

Discrimination manifests itself in unequal treatment of groups.

Other algorithms perform preprocessing on the training data.[11,14,19,26] These works can be seen as attempting to reconstruct the construct feature space and make decisions based on that hypothesized reality under the WAE assumption.

We turn now to Dwork et al.'s individual fairness definition:[10] two individuals who are similar should receive similar outcomes. Dwork et al. emphasize that determining whether two individuals are similar with respect to the task is critical and assume such a metric is given. In light of the formalization of the construct spaces, we note that the metric discussed by Dwork et al. is the distance in a combined construct space including both features and decisions. As described by Dwork et al., the metric is not known. We claim that in practice this lack of knowledge is resolved by the axiomatic assumption of either WYSIWYG or WAE, and since the focus is on individual fairness and not on groups, the WYSIWYG assumption is usually made between both feature and decision spaces.

TL;DR

Fairness **measure** choices encode assumptions about the **decision** spaces. Parity-focused notions (for example, disparate impact) assume WAE. Error rate balance assumes WYSIWYG.

Intervention **algorithm** choices encode assumptions about the **feature** spaces. Representational approaches assume WAE. In-processing and post-processing approaches assume WYSIWYG.

Fairness meta-analyses. Further examination of group fairness measures prompted the discovery of the mutual incompatibility of error rate balance (for both positive and negative classifications) and equality of per-group calibration (a measure indicating if a score is correctly predicting gradated outcomes). These constraints cannot be simultaneously achieved unless the classifier is perfect or the base rates per group are equal.[7,17] Since these measures naturally make a WYSIWYG assumption between the decision spaces, this impossibility result only holds under this axiom. In fact, the case under which it no longer holds—the base rates per group being equal—is one possible codification of the WAE axiom in the construct decision space.

TL;DR

Fairness impossibility results[7,17] hold under the WYSIWYG axiom between the decision spaces. These results do not hold under the WAE axiom between decision spaces.

Meta-analyses should make their axiomatic assumptions explicit and consider both measures and algorithms.

Discussion and Conclusion

Our main claim in this work is that discussions about fairness algorithms and measures should make explicit the implicit assumptions about the world being modeled. The focus by traditional data science techniques on the observed feature and decision spaces obscures these important axiomatic issues. The default assumption in these traditional data science and machine learning domains is the WYSIWYG assumption; the data is taken as given and fully representative of the implicit construct spaces. In this work, we highlight that this WYSIWYG assumption should be made purposefully and explicitly.

When considering fairness-aware algorithms applied to a specific domain, all assumptions are not equally reasonable. There is extensive social science literature demonstrating the existence of structural bias in criminal justice,[2] education,[18] and other fairness-critical domains. In these domains, it is not reasonable to make the WYSIWYG assumption. Data science practitioners must work with domain experts and those impacted by resulting decisions to understand what assumptions are reasonable in a given context before developing and deploying fair mechanisms; without this work, incorrect assumptions could lead to unfair mechanisms.

Additionally, our framework suggests ways in which the current discussion of fairness measures is misleading. First, group and individual notions of fairness reflect fundamentally different underlying goals and are not mechanisms toward the same outcome. Second, group notions of fairness differ based on their implicit axiomatic assumptions: mathematical incompatibilities should be viewed as a formal statement of this more philosophical difference. And finally, and perhaps most importantly, comparing definitions of fairness is incomplete without also discussing the deployed interventions: it is the combination of measure and algorithm that describes a fully specified worldview in which the system operates.

Acknowledgments. This research was funded in part by the NSF under grants IIS-1251049, CNS-1302688, IIS-1513651, IIS-1633724, and IIS-1633387. Thanks to the attendees at the Dagstuhl Workshop on Data Responsibly for their helpful comments, and to Cong Yu, Michael Hay, Nicholas Diakopoulos and Solon Barocas. We also thank Tionney Nix, Tosin Alliyu, Andrew Selbst, danah boyd, Karen Levy, Seda Gürses, Michael Ekstrand, Vivek Srikumar, and Hannah Sassaman and the community at Data & Society.

References

1. Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *Proceedings of the 35th Intern. Conf. on Machine Learning 80*. J. Dy and A. Krause, (Eds.). PMLR, (Stockholmsmässan, Stockholm Sweden, 2018), 60–69; http://proceedings.mlr.press/v80/agarwal18a.html
2. Alexander, M. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness.* The New Press, 2012.
3. Almlund, M., Duckworth, A., Heckman, J., and Kautz, T. *Personality psychology and economics.* Technical Report w16822. NBER Working Paper Series. National Bureau of Economic Research, Cambridge, MA, 2011.
4. Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. *ProPublica* (May 23, 2016).
5. Barocas, S. and Selbst, A. Big data's disparate impact. *California Law Review 104*, 671, (2016).
6. Calders, T. and Verwer, S. Three naïve Bayes approaches for discrimination-free classification. *Data Min Knowl Disc 21* (2010), 277–292.
7. Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data 5*, 2 (2017), 153–163.
8. Datta, A., Tschantz, M., and Datta, A. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. In *Proceedings on Privacy Enhancing Technologies 1* (2015), 92 – 112.
9. Duckworth, A., Peterson, C., Matthews, M., and Kelly, D. Grit: Perseverance and passion for long-term goals. *J. Personality and Social Psychology 92*, 6 (2007), 1087–1101.
10. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel. R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conf.* (2012), 214–226.
11. Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining*, (2015), 259–268.
12. Friedler, S., Scheidegger, C., and Venkatasubramanian, S. On the (im)possibility of fairness; *arXiv:1609.07236* (2016).
13. Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 2016, 3315–3323.
14. Kamiran, F. and Calders, T. Classifying without discriminating. In *Proceedings of the 2nd Intern. Conf. Computer, Control and Communication.* IEEE, (2009), 1–6.
15. Kamiran, F., Karim, A., and Zhang, X. Decision theory for discrimination-aware classification. *ICDM*, (2012), 924–929.
16. Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. Fairness-aware classifier with prejudice remover regularizer. *Machine Learning and Knowledge Discovery in Databases* (2012), 35–50.
17. Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of Innovations in Theoretical Computer Science*, (2017), 43:1–43:23.
18. Kozol, J. *The Shame of the Nation: The Restoration of Apartheid Schooling in America.* Broadway Books, 2006.
19. Madras, D., Creager, E., Pitassi, T., and Zemel, R. Learning adversarially fair and transferable representations. In *Proceedings of the 35th Intern. Conf. on Machine Learning 80*. J. Dy and A. Krause (Eds.), PMLR, (2018), 3384–3393; http://proceedings.mlr.press/v80/madras18a.html
20. Roemer, J. *Equality of Opportunity.* Harvard University Press, 1998.
21. Romei, A. and Ruggieri, S. A Multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* (Apr. 3, 2013), 1–57.
22. Ruggieri, S. Using t-closeness anonymity to control for nondiscrimination. *Transactions on Data Privacy 7* (2014), 99–129.
23. Yeom, S. and Tschantz, M. Discriminative but Not Discriminatory: A Comparison of Fairness Definitions under Different Worldviews. *CoRR* abs/1808.08619 (2018). arXiv:1808.08619 http://arxiv.org/abs/1808.08619
24. Zafar, M., Valera, I., Rodriguez, M., and Gummadi, K. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of WWW*, (2017), 1171–1180.
25. Zafar, M., Valera, I., Rogriguez, M., and Gummadi, K. 2017. Fairness constraints: Mechanisms for fair classification. *Artificial Intelligence and Statistics*, (2017), 962–970.
26. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *Proceedings of ICML*, (2013), 325–333.
27. Zliobaite, I. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery 31*, 4 (2017), 1060–1089.

Sorelle A. Friedler (sorelle@cs.haverford.edu) is an associate professor of computer science at Haverford College, Haverford, PA, USA.

Carlos Scheidegger (cscheid@email.arizona.edu) is an associate professor of computer science at the University of Arizona, Tucson, AZ, USA.

Suresh Venkatasubramanian (suresh@cs.utah.edu) is a professor of computer science at the University of Utah, Salt Lake City, UT, USA.