This is the author manuscript, before publisher editing. The citation for the published version is as follows: M. Altman et al., "A Harm-Reduction Framework for Algorithmic Fairness," IEEE Security & Privacy, vol. 16, no. 3, pp. 34-45; doi:10.1109/MSP.2018.2701149.

A Harm-Reduction Framework for Algorithmic Fairness[1]

by Micah Altman,[2] Alexandra Wood,[3] and Effy Vayena[4]

Abstract: In this article we recognize the profound effects that algorithmic decision-making can have on people's lives and propose a harm-reduction framework for algorithmic fairness. We argue that any evaluation of algorithmic fairness must take into account the foreseeable effects that algorithmic design, implementation, and use have on the well-being of individuals. We further demonstrate how counterfactual frameworks for causal inference developed in statistics and computer science can be used as the basis for defining and estimating the foreseeable effects of algorithmic decisions. Finally, we argue that certain patterns of foreseeable harms are unfair. An algorithmic decision is unfair if it imposes predictable harms on sets of individuals that are unconscionably disproportionate to the benefits these same decisions produce elsewhere. Also, an algorithmic decision is unfair when it is regressive, i.e., when members of disadvantaged groups pay a higher cost for the social benefits of that decision.

Keywords: data privacy, informational harm, fairness, accountability

---

[1] We describe contributions to the paper using a standard taxonomy (see Allen L, Scott J, Brand A, Hlava M, Altman M. Publishing: Credit where credit is due. Nature. 2014;508(7496):312-3). Altman provided the core formulation of the paper's goals and aims, and Altman and Wood led the writing of the original manuscript. All authors contributed to conceptualization through additional ideas and through commentary, review, editing, and revision. This material is based upon work supported by the Alfred P. Sloan Foundation and the National Science Foundation under Grant No. 1237235. The manuscript was prepared for the 2nd Annual Brussels Privacy Symposium, "AI Ethics: The Privacy Challenge," hosted by the Brussels Privacy Hub of the Vrije Universiteit Brussel and the Future of Privacy Forum in Brussels, Belgium, on November 6, 2017. The authors wish to thank their colleagues through the Harvard Privacy Tools Project for articulating ideas that underlie many of the conclusions drawn in this paper.

[2] escience@mit.edu. Dr. Micah Altman is Director of Research and Head/Scientist, Program on Information Science for the MIT Libraries, at the Massachusetts Institute of Technology. Dr. Altman is Chair of the National Digital Stewardship Alliance coordination committee; serves on the boards of directors for ORCID and iSolon; on the executive board for the American Political Science Association's section on Information Technology and Politics; the steering committee for the Data Preservation Alliance for Social Science; on the technical advisory boards of Force11 and the Qualitative Data Archive; and on the editorial boards of the *American Journal of Political Science*, *Social Science Computer Review*, *Journal of Information Technology and Politics* and Statistical Associates Publishers. Dr. Altman earned a PhD in Social Science from the California Institute of Technology, and conducted his postdoctoral research at Harvard University.

[3] awood@cyber.law.harvard.edu. Alexandra Wood is a Fellow at the Berkman Klein Center for Internet & Society at Harvard University and a Senior Researcher contributing to the Harvard Privacy Tools Project. Her research explores legal and regulatory frameworks for privacy and data protection in light of recent advances in privacy from fields such as computer science, social science, and law. She holds a law degree from George Washington University Law School, a master's degree in public policy from the University of Southern California, and a bachelor's degree in economics from Reed College.

[4] effy.vayena@hest.ethz.ch. Dr. Effy Vayena is Professor of Bioethics at the Swiss Federal Institute of Technology (ETH Zurich) and a member of the Swiss Academy of Medical Sciences. Her research focus is on the ethics of data uses, data governance, and specifically the implication of uses in the domain of heath. She holds degrees in History of Medicine and Bioethics.

I. Individual, group, and societal interests in control over information collected, sharing and use

Artificial intelligence and machine learning are increasingly applied to personal information and used to make decisions that affect the lives of individuals in ways large and small. Examples include algorithms used by online retailers to tailor prices to consumers based on estimates of their location and by automobile insurance companies to calculate premiums based on factors such as the length of a customer's commute (see [1], [2]). Law enforcement officers also use facial recognition algorithms to identify suspects appearing in footage from a crime scene, judges consult risk assessment algorithms on bail, sentencing, and parole decisions based on an individual's demographic characteristics and criminal history, and airport security screeners make use of algorithmically-determined risk assessment scores for airline passengers (see [1], [2]). There are countless other examples from consumer, employment, education, health care, credit, insurance, finance, criminal justice, and national security applications, with the development and adoption of algorithmic approaches to decision-making continuing to expand rapidly.

The potential for algorithmic decision-making to result in serious harm to individuals has been widely recognized in the literature, particularly within highly consequential contexts such as criminal justice, health care, education, and employment (see, e.g., [1] - [4]). Whether entirely or partially automated, algorithmic approaches to collecting, analyzing, classifying, and making decisions with respect to personal data can profoundly affect the wellbeing of individuals, groups, and society. The use of such approaches is intended to enhance predictions and recommendations in ways that are substantially beneficial to groups of people through improvements in areas such as public safety, health outcomes, or efficiency. At the same time, applications of algorithmic decision-making challenge individual, group, and societal interests regarding control over information collection, sharing, and use, as well as notions of privacy, equity, fairness, and autonomy (see [5]). In particular, they have the potential to create unintended and even unforeseen harmful effects, including contributions to systematic inequality of opportunity for socially disadvantaged individuals and communities.

Yet current approaches to regulating algorithmic classification and decision-making elide an explicit analysis of harm—despite the emphasis on harm under analogous regulatory frameworks for uses of personal information, such as the ethical framework for the protection of human subjects participating in research. Existing legal and ethical frameworks have been developed to prevent or provide redress for some of the same types of harms resulting from automated decision-making. The ethical norms embodied in the law point to the broader responsibilities of the architects and users of algorithms that make decisions that affect the lives of individuals.

Although much attention has been drawn to the immediate harms of criminal use of information stemming from data breaches, harms that stem from non-criminal use of personal information are not very well understood. An understanding of harm from algorithmic classification and decision-making can be informed by existing legal and ethical frameworks. Harm is a central concept in law and ethics, from the body of common law tort theory to the injury-in-fact requirement of the standing doctrine, and from the elements of criminal offenses to the ethical principles underlying human subjects protection regulation (see [6]). Notably, the Belmont Report developed by the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research in 1978 encapsulates fundamental ethical

principles that "serve as a basic justification for the many particular ethical prescriptions and evaluations of human actions" [7]. Among the fundamental principles is "respect for persons," which requires protecting persons from harm following the principle that "[t]he extent of protection afforded should depend upon the risk of harm and the likelihood of benefit" [7]. In recognition of the principle of respect for persons, ethical frameworks encompass broad notions of harm, guiding actors to consider risks of psychological, physical, legal, social, and economic harm [7]. The Menlo Report includes a restatement of these principles in the context of information and communication technology research (see [8]).

Although some regulatory approaches require an ex-ante review of the types of information that may permissibly be used as inputs to an algorithmic decision or transparency about the process that is used, to our knowledge, few, if any, approaches require a systematic ex-ante review of potential harms to individuals. More rigorous assessment of harm to individuals is likely to play an increasingly central role in regulatory and ethical review of algorithmic decision-making. Notably, the new European General Data Protection Regulation (GDPR) explicitly recognizes the importance of taking harm into consideration when automated decision-making is used, requiring data controllers to provide data subjects not only with "meaningful information about the logic involved" in the automated process but also information about "the significance and the envisaged consequences of such processing for the data subject" [9].

We argue, generally, that explicit analysis of algorithmic fairness, based on counterfactual analysis, should be incorporated into algorithm design. We illustrate this approach through an analysis of automated risk assessment in the criminal justice context. Specifically we identify four elements that should be incorporated into any analysis of algorithmic fairness:

- Identification of the major choices in algorithmic design, implementation and application that have the potential to predictably and substantially affect well-being;
- Assessment, using counterfactual causal estimation, of the effects of these decisions on the well-being of individuals;
- Measurement of well-being broadly, to include lifetime wealth, health, longevity, subjective life-satisfaction, and the ability to make ethically relevant choices; and
- Recognition of algorithmic unfairness as choices in algorithmic design, implementation, and application that have disproportionate effects on members of different groups.

II. Using COMPAS to understand harm: Algorithmic discrimination in predictions of an individual's risk of recidivism or failure to appear in court

A particularly compelling real-world example of the potential for algorithmic discrimination and harm is the use of risk assessment scores within the criminal justice system. Judges are increasingly using automated decision support software, such as the Northpointe Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) risk assessment algorithm, to predict an individual's risk of recidivism or failure to appear in court based on various factors, including his or her criminal history and socioeconomic characteristics. COMPAS was developed with the aim of informing decisions across a

range of stages in the criminal justice process, including pretrial, jail, probation, prison, parole and community corrections decisions. It has been deployed for use in Florida, Michigan, New Mexico, Wisconsin, Wyoming, among other states.

A 2016 ProPublica analysis of the use of COMPAS risk scores in Broward County, Florida, uncovered evidence of racial bias, finding that when the effect of race, age, and gender was isolated from criminal history and recidivism risk, "[b]lack defendants were still 77 percent more likely to be pegged as at higher risk of committing a future violent crime and 45 percent more likely to be predicted to commit a future crime of any kind" [10]. They also found that COMPAS "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" [10]. In addition, COMPAS "proved remarkably unreliable in forecasting violent crime: Only 20 percent of the people predicted to commit violent crimes actually went on to do so" [10].

Implicitly underlying criticisms of the use of automated risk assessments like COMPAS is the concern that they have the potential to make individuals worse-off. Adverse sentencing decisions have long-ranging consequences for individuals including decreased quality of life over a substantial period of time and reduced lifetime job prospects and earnings due to one's criminal conviction. Similarly, adverse bail decisions can result in harms to low-risk defendants such as loss of freedom, damage to familial relationships and family members, and losses in future employment and earnings.

It is also important to recognize that the adverse effects of algorithmic decisions may reach beyond the individual directly subject to that decision. Harmful effects can also reach family members, communities, and society at large. Research has shown that children of inmates exhibit higher rates of mental health problems, behavioral problems, financial insecurity, poor school performance, and infant mortality (see [11]). Incarceration rates also vary substantially based on race and class, exacerbating societal issues of inequality. The literature also cites the disproportionate effects that mass incarceration have had on African American communities by "damaging social networks, distorting social norms, and destroying social citizenship" [12]. A group may be harmed disproportionately to the social benefit, and this is especially problematic if the group is a historically disadvantaged group. In addition, if the harms are not connected to choices made by the individuals, it may be clear that the absolute cost to those individuals is too high.

These outcomes are ethically relevant because they are persistent and substantially affect multiple aspects related to an individual's life satisfaction. Where the likelihood of adverse decisions is predictably affected by choices in the design, implementation, and application of algorithms, the underlying algorithmic choices demand ethical analysis. An ethical grounding for evaluations of COMPAS and other algorithmic decisions can be found in the life course approach (see, e.g., [11], [13]). This approach enables a comprehensive analysis of how a socially defined event, such as incarceration or pre-trial detention, influences an individual's life trajectory, and its quality. Applications of this framework recognize the critical ethically relevant fact that the consequences of some harmful events are not only realized immediately subsequent to the decision but also have potential long-term effects throughout the course of an individual's life.

Different fields of study, such as economics or public health, focus on different aspects of the life course, in part because these fields focus on different potential interventions. However, there is no reason to believe that a single aspect of the life course, such as health, is the only measure relevant to an ethical review. When analyzing fairness, one should measure all of the aspects of life that are widely recognized within social science and health fields as fundamental for well-being. Specifically, the literature identifies five key measures for a life course analysis: wealth, lifespan, health, subjective life-satisfaction, and the ability to make substantial choices about one's life (sometimes referred to as "capability").[5]

Life course analysis can be applied to understand how the consequences of incarceration can be greater for some individuals than for others. For example, life course research has shown that incarcerating an individual during the critical period of transition into adulthood drastically alters the course on an individual's life, putting that individual "off-time" with his or her peers, and decreasing the likelihood that he or she will eventually obtain a college degree, find stable employment, form a family, or participate in civic life (see [11]). Certain categories of individuals may be especially vulnerable for other reasons as well. For example, research has found that some groups, such as the young, small of stature, and mentally disabled are more likely to become victims of sexual assault during incarceration (see [16]).

III. Applying life course analysis to critical choices in the design and application of the COMPAS recidivism risk score

When can an algorithmic choice be expected to cause a change in a subject's well-being? Although complex to estimate empirically, the answer can framed as a causal inference problem, using the potential-outcomes framework broadly recognized in statistics, social sciences, health science, and computer science (see, e.g., [17], [18]). Further, it is likely that a causal-counterfactual mode of analysis is the only one that can yield empirically reliable inferences based on heterogeneous, complex, observational data.[6] An observation from the technical literature is that big data inference can be considered reliable if and only structural causal model information is included. This implies there is a confluence between the requirements of fairness and those of reliable inference.

Viewed through the lens of the potential-outcomes framework, intuitively, an algorithmic decision causes harm to an individual when the expected outcome for that person given the decision is worse than the expected outcome for that person absent the decision. Applying this framework to the evaluation of real-world risk assessment algorithms used in the criminal justice system, we can identify the relevant counterfactuals and find that they can be shown to yield concrete measures of harm to individuals. When developing and applying such an algorithm, there are at least four phases where key decisions can be made. As illustrated in Figure 1, these phases encompass choices related to the background or training data to include as inputs to the algorithm, the design and implementation of the algorithm, the application of the algorithm to a particular individual, and the use of the score in a given sentencing decision.

[5] For a survey of life-satisfaction models and the measures commonly used, see [14]. For a discussion of the capability approach to individual and social analysis, see [15].

[6] An observation from the technical literature is that big data inference can be considered reliable if and only structural causal model information is included. This implies there is a confluence between the requirements of fairness and those of reliable inference. See [19].

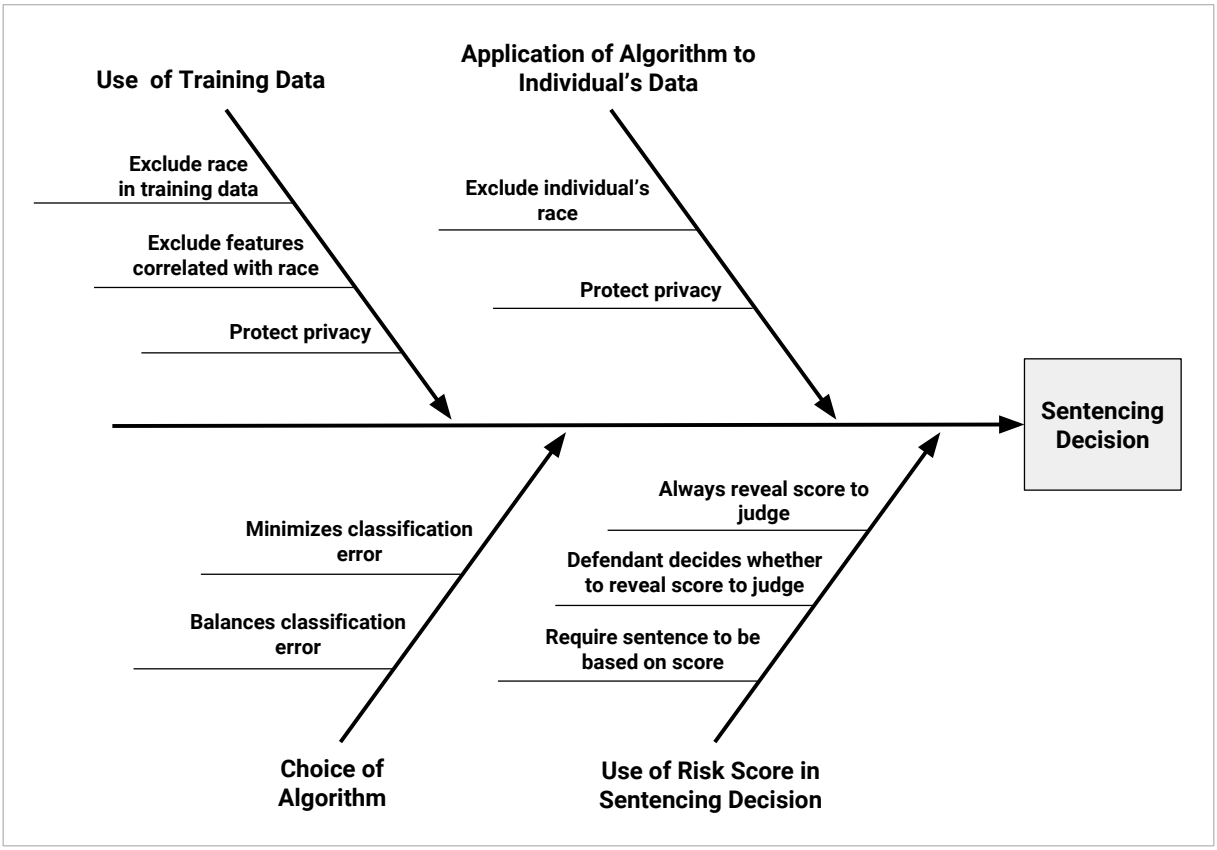

Figure 1. Critical choices in the design and application of the COMPAS algorithm.

To provide a detailed illustration, we focus on a series of counterfactuals relevant to classification algorithms used in criminal sentencing. Specifically, we analyze counterfactuals for potential interventions such as excluding racial data as inputs to the algorithm, running analyses on the data using differential privacy, substituting human judgment for an algorithm's predictions, and utilizing an algorithm that balances classification error rates. The analysis identifies the relevant choice that was made, such as the inclusion of the individual's information in an instance of data collection, storage, or computation. It also considers the available alternatives by way of, for example, a ceteris paribus analysis (that is, every fact remains the same except for that affected by a single choice), the comparison of the effects resulting from the use of one protective measure vs. the status quo or the use of another measure, and an opt-out rule, in which each individual has the right to opt out of an analysis. These examples are used to illustrate how harm can be measured in various ways based on the effect on an individual's life trajectory in each use case. Example harms range from dollar amounts, to health outcomes—physical or emotional—and other quality of life measures, and from educational attainment to effects on one's autonomy and liberty.

Counterfactual #1: What if the decision were not based on race? Exclusion of racial data from inputs to the algorithm

For the first counterfactual, we explore the effect of excluding information about race from the decision-making algorithm. This could be interpreted as excluding racial information directly, or, perhaps, excluding all information correlated with race from the algorithm. In these counterfactuals, the relevant choice could have been made at either the collection or analysis stage, depending on which point the decision to include or exclude certain data as inputs to the algorithm was reached.

Consider what happens when excluding protected attributes such as racial data from the model. Dwork et al. [20] and others have widely concluded that fairness through blindness fails because of redundant encodings. Because other attributes act as proxies for the protected attributes, the expected difference between the factual and counterfactual scenarios is minimal. In fact, the impact of the analysis may be identical under the factual and counterfactual scenarios, as illustrated in Figure 2 below.

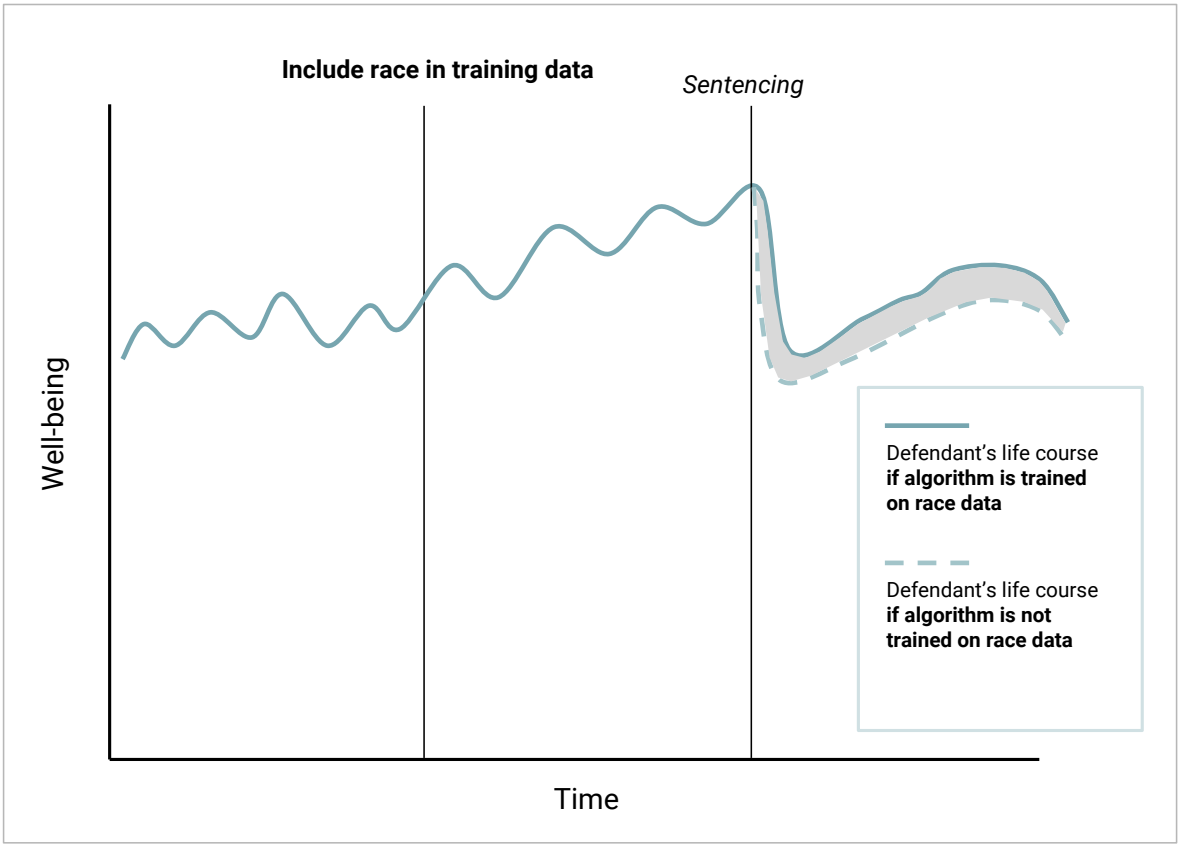

Figure 2. Well-being over a non-recidivist minority defendant's life course with the adverse algorithmic decision, and with the counterfactual algorithmic decision of excluding protected attributes such as racial data as inputs to the analysis.

Because removing the protected attributes will not prevent other attributes correlated with race from having an impact on the analysis, one may be tempted instead to remove all of the measures that have a

substantial correlation with race. However, it may be the case that all of the measures that are likely to predict recidivism have a substantial correlation with race. Removing all such measures would leave no data, or very limited data, with which to make predictions. In addition, treating individuals in different groups fairly may require taking protected features into account (see [20]). Removing protected attributes or attributed correlated with protected attributes, or prohibiting the release or analysis of such data, may also make it difficult to study the effect of discrimination based on race in criminal sentencing (see [21]).

Counterfactual #2: What if participants' privacy were protected? Using differential privacy as a control on the analysis

Second, we explore the counterfactual in which measures are taken to protect the privacy of individuals when using their personal information. Specifically, we look at the example of the use of a formal privacy model, differential privacy, as a control on the analysis. Differential privacy is a mathematical guarantee of privacy protection in the context of statistical and machine learning analysis. It provides a quantifiable measure of the excess privacy risk any individual contributor may incur due to her participation in an analysis, compared with not participating. What can be learned about an individual contributor from a differentially private analysis is (essentially) limited to what could be learned about that individual from everyone else's data without his or her own data being included in the analysis.

In this counterfactual scenario, the relevant decision is made at the data analysis stage. More concretely, consider the scenario in which differential privacy is used in a machine learning model for predicting a defendant's recidivism risk. A machine learning algorithm that satisfies differential privacy would be trained on historical data about defendants, their criminal histories, their demographic information, and their actual incidences of recidivism. When differential privacy is used, a model based on the training data would reveal very little about an individual in the training data, such as information about his or her criminal history, but would enable the relationship between demographic characteristics, criminal history, and recidivism to be determined at the population level. The use of differential privacy would not prevent the machine learning algorithm from identifying and applying a trend in the training data such as defendants from a particular minority group being associated with a higher than average rate of recidivism.

In order to make a decision with respect to an individual, that individual's information must be used. When the differentially private machine learning model is applied to a defendant's information in order to make a prediction about the defendant's risk of recidivism, the individual's information will be used and the predictive power of certain protected features, such as race, will continue to hold. For instance, if the model identifies a positive correlation between the defendant's race and recidivism risk, the model may classify the defendant as a high recidivism risk, just as it would in the factual scenario.

Figure 3 illustrates this counterfactual scenario. It shows how an individual's life course is expected to be largely unaffected by the use of differential privacy. Differential privacy provides protection against learning about an individual's attributes from the machine learning model in cases in which the individual's information is included in the training data and the model is not applied to make predictions about that individual. However, it does not provide protection in cases in which the model is applied to

the individual's information in order to make predictions about the individual, which in many cases will lead to much greater effects on that individual's life course than the former.

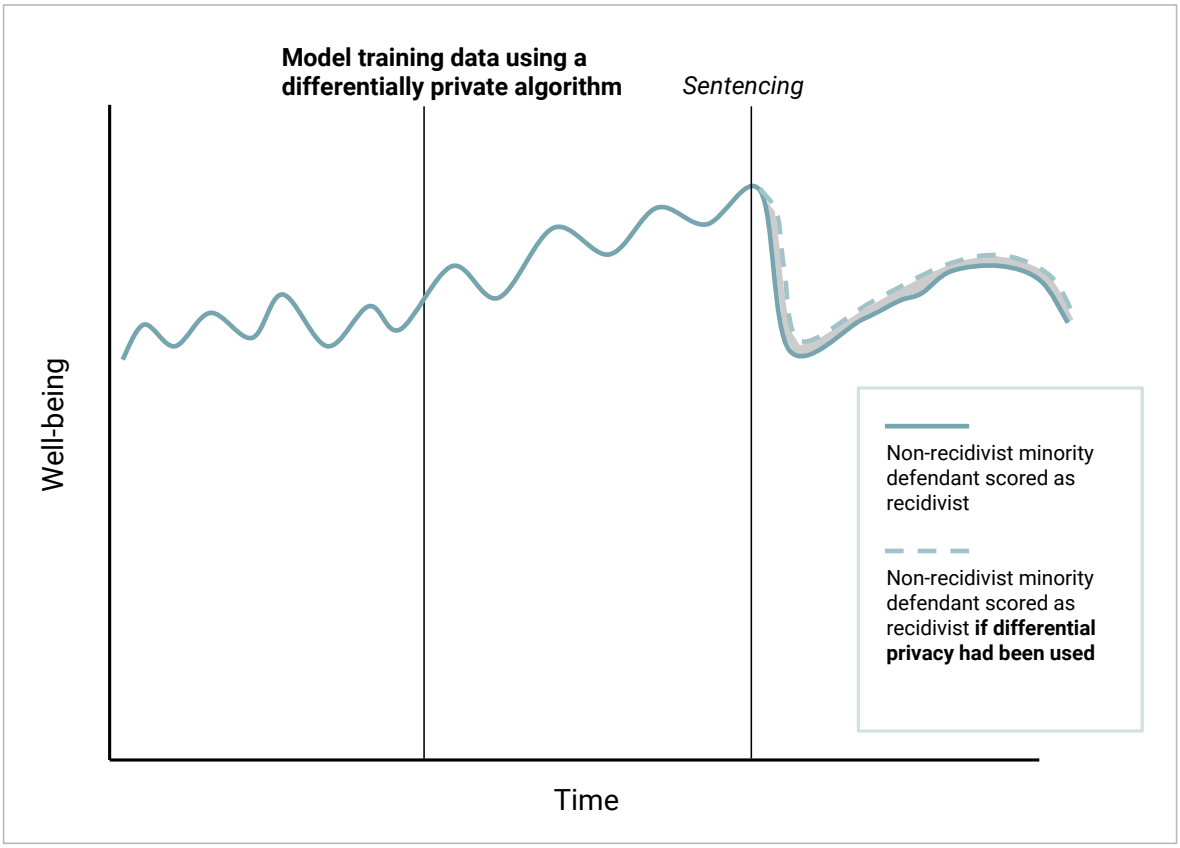

Figure 3. Well-being over a non-recidivist minority defendant's life course with the adverse algorithmic decision, and with the counterfactual algorithmic decision of using differential privacy as a control on the analysis.

Therefore, differential privacy alone will not protect the defendant from all harms—or even the most relevant harms—from application of the model. This counterfactual illustrates how privacy may in some cases be orthogonal to other data use-related harms. Privacy controls can protect the privacy of individuals in an analysis, while still failing to protect them from non-privacy-related harms resulting from the analysis. In many cases, other harms, such as discrimination, will far exceed the privacy-related harms that are mitigated by the privacy intervention.

A related counterfactual is one in which a defendant is given the opportunity to choose whether to reveal his or her risk score to a judge prior to sentencing. This counterfactual leads to a signaling equilibrium, in which scores predicting a high likelihood of recidivism are inferred (albeit with some uncertainty) from a choice not to reveal the score. This is an example of a more general observation that the potential for signaling equilibria undercuts the assumption that allowing individuals to choose to withhold or reveal personal information systematically protects them (see [22]).

Counterfactual #3: What if an algorithm were not used at all? Substituting human judgment for an algorithm's predictions

Third, we explore the counterfactual difference between basing a sentencing decision on an algorithmic prediction and basing a sentencing decision on human judgment. When making a sentencing decision, a judge identifies rational correlations between individual attributes and outcomes in order to make a prediction about a defendant's recidivism risk, resembling a machine learning algorithm in key ways. This process, however, is subject to the limits of human decision-making, in addition to the limits of computational decision-making. A judge's predictions about a given defendant's risks of recidivism are inherently imperfect, whether due to imperfect information, imperfect mental models, or limited experience. For this counterfactual, the literature on sources of error in human decision-making, empirical research estimating the error rates of judge-made decisions, and studies comparing the performance of human decisions and algorithmic predictions can be instructive.

Kleinberg et al. [23], for example, demonstrate the potential for algorithms to provide improved predictions compared to judge-made decisions. The authors evaluate a machine learning algorithm for making predictions about judges' pre-trial detention decisions and estimate that decisions based on the algorithm's predictions could reduce the crime rate by 24.7% while holding the jail detention rate constant, or reduce the jail detention rate from 26.4% to 15.4% holding the crime rate constant (see [23]). While the focus of this research is pre-trial detention decisions, where what is being predicted is flight risk rather than recidivism risk, it is closely analogous to sentencing decisions and suggests that algorithmic approaches have the potential to produce large gains over human decisions. One possible explanation for the counterfactual difference between the algorithmic and human decisions is the reliance by judges on unobserved attributes that are not good predictors of flight risk. The authors note that "[w]hatever these unobserved variables are that cause judges to deviate from the predictions—whether internal states, such as mood, or specific features of the case that are salient and over-weighted, such as the defendant's appearance—they are not a source of private information so much as a source of mis-prediction. The unobservables create noise, not signal" [23] (citations omitted).

Figure 4 illustrates the expected counterfactual difference between algorithmic and human decisions in the context of sentencing decisions, reflecting the results of research demonstrating that algorithmic decisions have the potential to better isolate the signal from the various attributes used as the basis for predictions of an individual's recidivism risk. This diagram accordingly shows how the adverse impact on an individual's life course can potentially be much greater when the decision is made by a human and not informed by an algorithm's prediction. When comparing the decisions made by multiple judges to decisions informed by a single algorithm, not only does the expected value change, but the variance grows as well. Greater variance is undesirable, ceteris paribus, because individuals tend to be risk averse, especially when the risks are large.

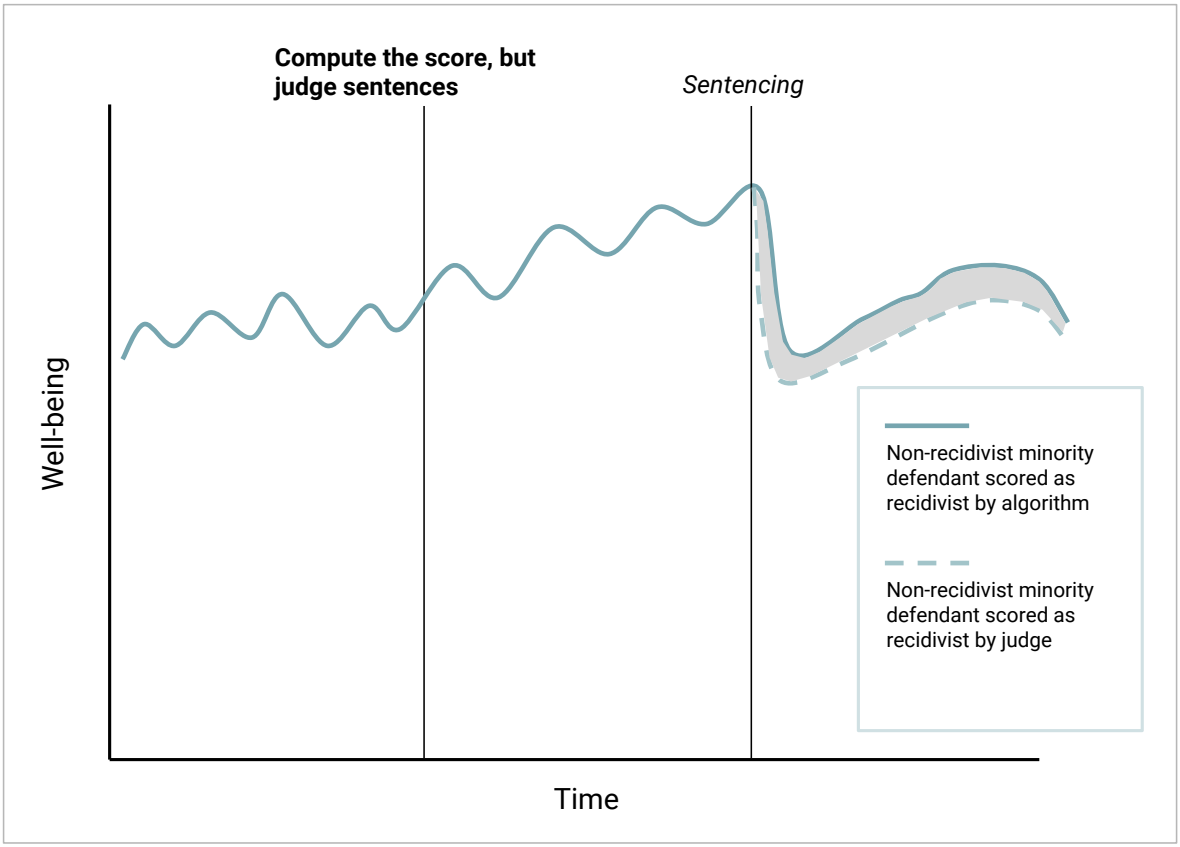

Figure 4. Well-being over a non-recidivist minority defendant's life course with the adverse algorithmic decision, and with the counterfactual human decision.

Counterfactual #4: What if a better algorithm were used? Comparing outcomes with those of a fairer or more predictive model

For this counterfactual, we examine the difference between outcomes under the algorithm being evaluated and a hypothetical model that is a stronger predictor of recidivism risk. This comparison depends on the metric used to assess whether an algorithm is better than another. For instance, an algorithm may be shown to be better because it provides strictly lower error rates for all individuals or Pareto improvements in error rates for historically disadvantaged populations. There are also practical challenges to empirical evaluations of error rates associated with applications of algorithm. For example, the analysis may require examining the sentencing decisions made in jurisdictions where judges are randomly assigned to cases. In addition, as mentioned above, assessing the fairness of an algorithm is challenging due to inherent tradeoffs between different fairness metrics. Another important consideration is that an algorithm used to guide sentencing decisions that is determined to be fairer or more accurate in its predictions will still lead to harms to the individuals who are incarcerated as a result.

For one example illustrating this counterfactual, consider again the Kleinberg et al. [23] machine learning model for pre-trial detention predictions. Their initial model did not include race or ethnicity data as inputs, although the authors note that "it is possible the algorithm winds up using these factors

distribution inadvertently—if other predictors are correlated with race or ethnicity" [23]. They explore the effects of applying various fairness constraints to this model and demonstrate that such constraints can "ensure no increase in racial disparities in the jail with very little impact on the algorithm's performance in reducing crime" or "reduce the share of the jail population that is minority—that is, reduce racial disparities within the current criminal justice system—while simultaneously reducing crime rates relative to the judge" [23].

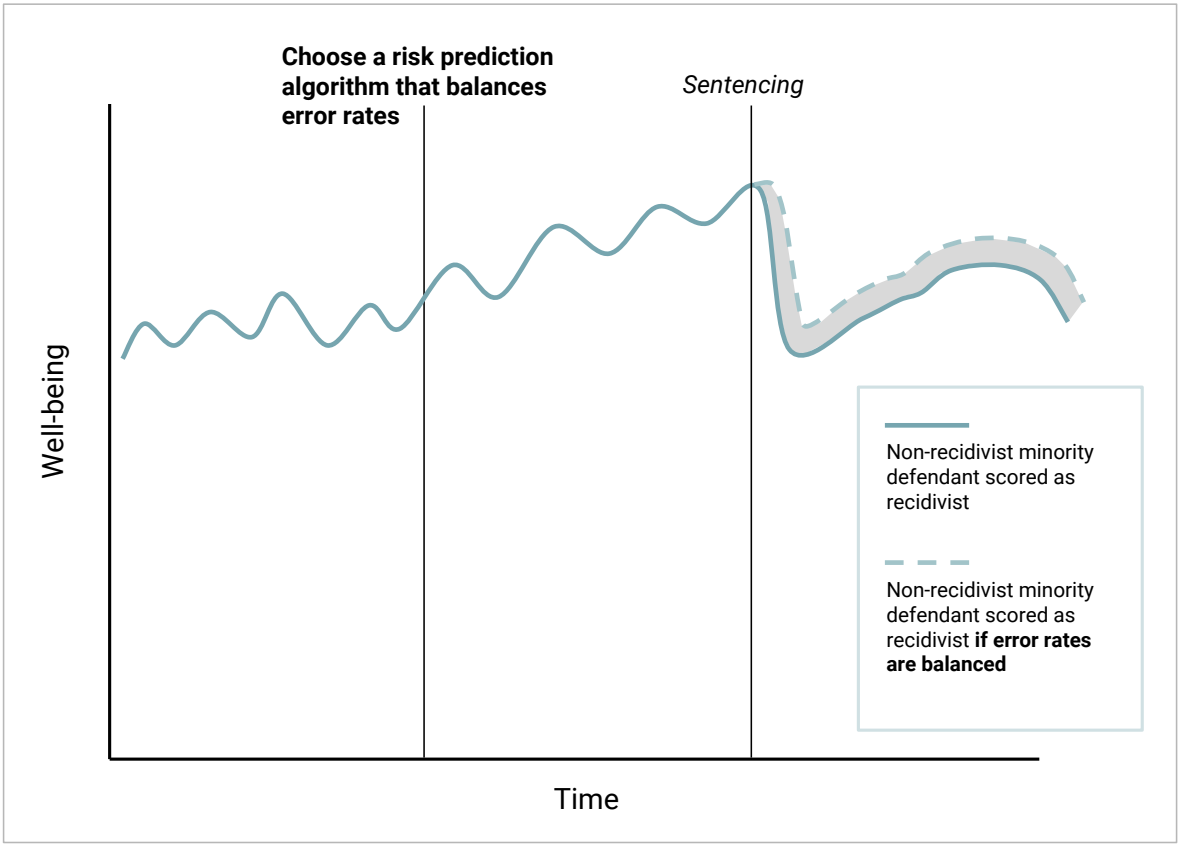

Figure 5. Well-being over a non-recidivist minority defendant's life course with the adverse algorithmic decision, and with the counterfactual of an algorithmic decision based on an algorithm that balances error rates.

Under this counterfactual, some individuals can be expected to benefit from the fairer algorithm. Some of the non-recidivist defendants misclassified as recidivists by the original algorithm would be classified correctly by the algorithm, leading to a reduction in harm. Although some individuals would remain misclassified by the algorithm and thus unimproved by its use, in the aggregate one can expect the overall expected level of harm to decrease. However, as we discuss in Section IV, such improvements alone do not necessarily ensure fairness, as the distribution of harms across groups could nevertheless be unfair.

The analyses of these counterfactuals illustrate several broader observations regarding gaps in traditional approaches. They suggest that certain types of protections, such as prohibiting the inclusion of sensitive characteristics as inputs to an algorithm or applying privacy measures such as de-identification or

differential privacy to protect information about individuals, are largely irrelevant when seeking to address fairness concerns. They also suggest that other protections, such as prohibiting discriminatory intent or requiring transparency of an algorithm's code, inputs, or logic, are insufficient to ensure fairness.

IV. Using the distribution of harms to evaluate fairness

The counterfactuals discussed above describe how choices affect the types and frequencies of errors made by algorithms and the potential harm to individuals subject to them. While an analysis of harm is an important component of determining whether an algorithm is fair, it is important to note that the existence of harm alone does not directly imply unfairness. Even fair algorithms can be expected to cause some peoples' lives to go worse. In order to determine whether an algorithm is fair, one must evaluate the distribution of harms across the population. This following discussion describes an approach to evaluating fairness by assessing the distribution of potential impacts under a particular counterfactual.

Surprisingly, algorithms that appear to be non-discriminatory can lead to imbalances in the distribution of costs and benefits for different groups to which they are applied. This can happen because the outcome depends not only on the algorithm itself, but also on a variety of empirical factors. Among these empirical factors are, primarily, the distribution of groups in the population; the types of classification made by the algorithm; the expected cost and benefit of each possible decision, based on that classification, for members of each group; and the distribution of errors made by the algorithm, with respect to each decision and each group.

This can be illustrated using the model of recidivism decision-making outlined in Section III. To simplify the analysis, we consider groups of people and types of classification in two categories, respectively: two groups (minority and non-minority); and two types of classification (non-recidivist or recidivist). Figure 6 below shows how these factors determine the distribution of outcomes, in the form of a decision tree.

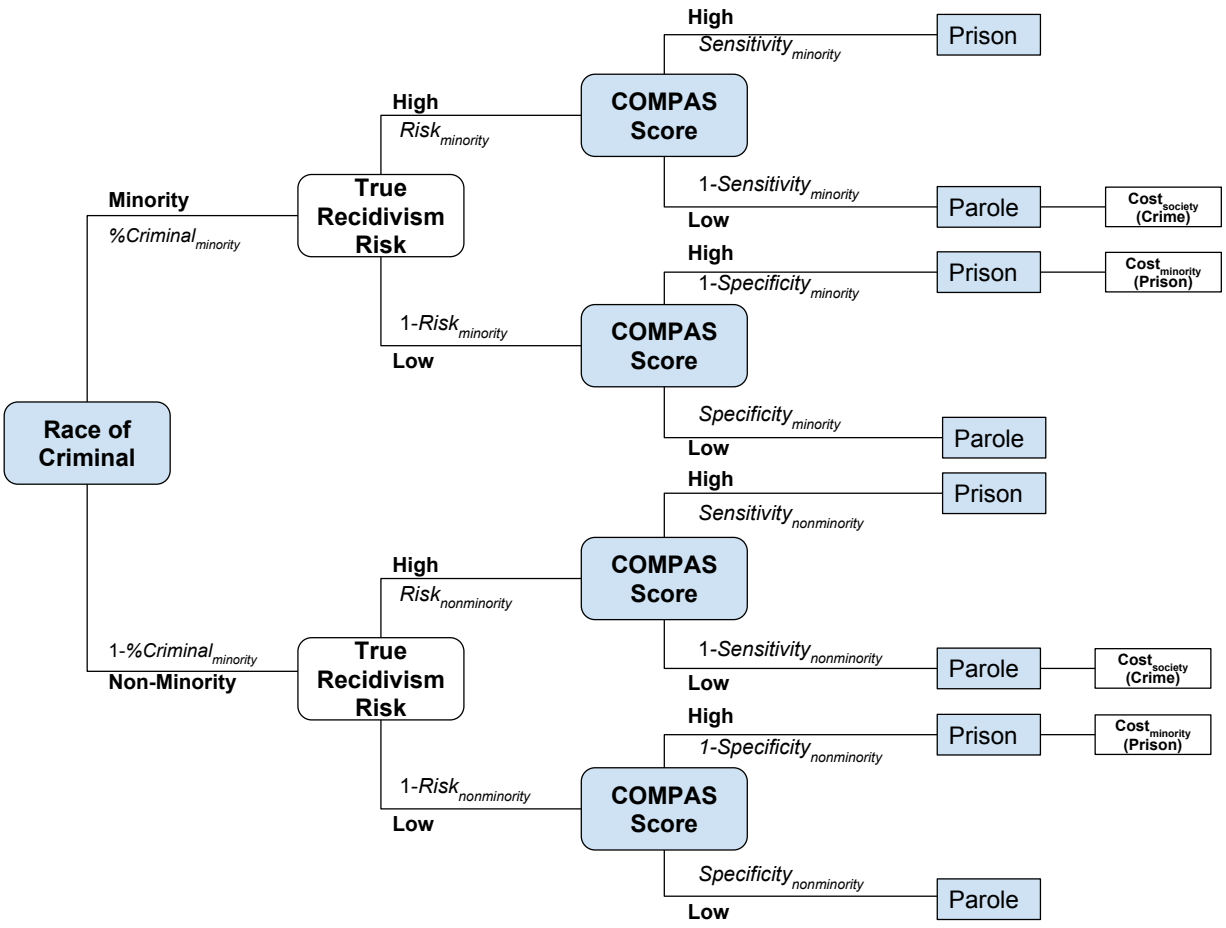

Figure 6. Decision tree showing how four factors lead to a distribution of outcomes. Decision factors are shown on the left, with rounded corners, and outcomes are shown on the right. Unshaded shapes indicate that the factor or outcome cannot be directly observed. For simplicity, only the costs associated with errors are shown.

From left to right, the diagram includes the race of the subject being scored, the true risk that the subject will recidivate (which is not known at the time of the decision), the COMPAS score assigned to the subject, the outcome resulting from that score, and the cost of that outcome. For simplicity of explanation in this section, we assume that all of the algorithmic design decisions have already been made, that parole decisions are directly determined by score, and that only incorrect decisions impose costs.

The formulation of this decision tree builds upon the ideas presented in previous sections. The *costs* that appear on the right-hand side of the decision tree can be estimated using the life course analysis discussed in Section II. The algorithmic decisions decisions discussed in Section III will influence the error rates of the algorithm (*sensitivity* and *specificity*, below), as well as the likelihood of a *prison* or *parole* decision based on a particular score. For example, under the third counterfactual discussed above, scores would influence the outcomes (i.e., change the likelihood of the outcome) but not be associated with a single deterministic outcome. Under the fourth counterfactual, the *sensitivity* and *specificity* rates would be improved and equalized across all groups.

A nominally fair algorithm will always yield an even distribution of costs across groups when each of the probabilities, percentages, and costs above are equal for each group. In other words, fair outcomes are guaranteed, if the criminal population is 50% minority, the risk of recidivism does not truly depend on race, the algorithm's error rate is identical for each member of the group, and the cost of mistakes is identical.

In almost all actual cases, however, some of the factors above will differ between groups, yielding differences in the distributions of outcomes. These distributions must be evaluated for fairness. In particular, there are three general conditions in which procedurally "fair" algorithms can yield unfair outcomes.

First, race blindness does not guarantee that the decisions will be equally accurate for each race. In other words the *sensitivity* and *specificity* rates (also known as the true positive and true negative rates) for the algorithm may differ based on group membership. Further, whenever the sensitivity and specificity of the algorithm have not been explicitly estimated for each group, they are very likely to differ—especially if, as explained in Section III, race is not an input to the algorithmic training. Worse, it is provably impossible to balance these error rates in practice [23].

Second, balancing accuracy rates—even in the idealized case—would not guarantee that mistaken decisions are equally distributed across the minority and non-minority groups. Algorithmic choices could yield substantially different distributions of outcomes for members of each group because of differences in the *prior* distribution of crime or true recidivism risk. This problem is commonly known as the "base rate fallacy" or "prosecutor's fallacy." For example, suppose, for the sake of argument, that the COMPAS algorithm were 99% accurate across the board, e.g., misclassifying only 1 out of 100 recidivists as low risk (and vice versa), regardless of race. Even in this case, the probability that a minority defendant is incorrectly sentenced could be much higher than the probability that a non-minority defendant is incorrectly sentenced—if the true (but hidden) recidivism risk for minority defendants is lower. Further, even if true recidivism risks were identical, the aggregate costs of incorrect decisions would be higher for minority groups, if minorities made up a higher proportion of the population of defendants.

Third, balancing the distribution of mistaken decisions would not guarantee that the distribution of costs is equal for members of each group, nor that the distribution of costs will be fair. The social-scientific analysis of life courses (discussed in Section II) demonstrates that the same event (such as parole) may have predictably different consequences for members of different groups. As noted in Section II, incarceration often has a larger long-term negative impact on adolescents than adults, by significantly reducing the likelihood that individuals in the former category will ever attain a degree, find stable employment, and form a family. Individual costs of incarceration may also vary by race and ethnicity. For example, one study has found that the negative effect of a criminal record on the likelihood of receiving a callback from a prospective employer is 40% greater for African-American job candidates than it is for white candidates [24].

In summary, under most realistic conditions the pattern of harm produced by mistaken algorithmic decisions will be unevenly distributed among groups. These patterns will depend on the algorithmic

design choices discussed in Section III; however, no practical algorithmic design choice is outcome-neutral. Thus, algorithm designers must choose, implicitly or explicitly, which types of errors are most important, and which groups should be classified more accurately, in order to yield a preferred distribution of harms and benefits. Algorithmic design choices are therefore to be considered ethically relevant choices.

Further, when algorithms are used in legal and government processes, there is quite frequently a social-choice problem that is being implicitly "solved." For example, the implied social benefits of using a recidivism risk score are better crime prevention through deterrence and incapacitation, which improves the lives of individuals who would otherwise become victims of crime. Thus the goal of the implied social choice problem is to balance the benefits to such individuals against the harms to the individuals scored and sentenced. In such an analysis, it is relevant to consider whether a decision to incarcerate an individual will harm that individual while also likely providing much greater benefits to others, and then assess how these harms and benefits are distributed across groups.

Although a full discussion of outcome fairness is beyond the scope of this article, we argue that a key question of fairness has been neglected in analyses of the use of recidivism risk scores in criminal sentencing. Namely, algorithmic decisions result in member of minority groups paying higher individual costs than members of non-minority groups, in proportion to the societal benefit gained.

When considering the fairness of algorithmic design decisions, we offer several recommendations. We recommend that the principle of progressive burden sharing be employed in cases in which the individual costs imposed by the use of an algorithmic decision are high and likely to be unequally distributed across different populations. Where, as is generally the case, the costs of algorithmic errors cannot be equalized across groups, algorithmic design choices should be considered unfair if they require members of less-privileged groups to pay higher costs (whether because of the frequency or severity of those errors) than members of more-privileged groups for the same types of algorithmic mistakes. Further, one should avoid algorithmic design choices that impose costs to some individuals that are highly disproportionate to the expected social gain. In particular, design choices that predictably, catastrophically, and persistently reduce the well-being of individuals who are members of a known class should be avoided. In addition, we recommend that algorithmic designers consider the harm incurred by all members of a group, in aggregate. For example, group punishment—meaning a choice that predictably, substantially, and persistently reduces the aggregate well-being of an entire historically disadvantaged class of individuals—is unfair and should be avoided in algorithmic design.

V. Recommendations: Accountability for algorithmic decisions

In this article we recognize the profound effects that algorithmic decision-making can have on people's lives and proposes a harm-reduction framework for algorithmic fairness. We argue that any evaluation of algorithmic fairness must take into account the foreseeable effects that algorithmic design, implementation, and use have on the well-being of individuals. We further demonstrate how counterfactual frameworks for causal inference developed in statistics and computer science can be used as the basis for defining and estimating the foreseeable effects of algorithmic decisions. Finally, we argue that certain patterns of foreseeable harms are unfair. An algorithmic decision is unfair if it imposes

predictable harms on sets of individuals that are unconscionably disproportionate to the benefits these same decisions produce elsewhere. Also, an algorithmic decision is unfair when it is regressive, i.e., when members of disadvantaged groups pay a higher cost for the social benefits of that decision.

An application of this framework to the COMPAS algorithm suggests that some common procedural interventions do not reliably improve algorithmic fairness. In particular, protecting informational privacy (e.g. through redaction, k-anonymity, or differential privacy), excluding information about protected attributes from a decision, and equalizing classification rates across protected groups do not generally increase algorithmic fairness. Further, under some conditions, these interventions may reduce fairness. Moreover, since accountability requires the ability to evaluate the effect of complex algorithmic decisions on the outcome in hypothetical situations, neither open data nor open source code is necessarily sufficient as an intervention [25].

In general, we find that because common procedural controls may have counterintuitive effects, there is an ethical responsibility to design, implement, and apply them in a manner that recognizes the potential for algorithmic choices to impose disproportionate harms on potentially vulnerable groups. To manage the potential for harm, we offer three practical recommendations.

First, we recommend that algorithms be designed and implemented in a way that facilitates the evaluation of how the harms from algorithmic decisions are distributed across groups. Section IV describes the factors that must be measured directly or estimated in order to facilitate such an analysis. These factors will generally include the relevant characteristics of individuals processed by the algorithm, the estimated ex-ante error rates for the algorithm when applied to individuals conditioned on different characteristics, and the distribution of actual algorithmic decisions for individuals with those characteristics.

Second, we recommend that explicit counterfactual causal analysis be used to predict the consequences of major algorithmic design decisions. As detailed in Section III, this analysis should evaluate the consequences of decisions with respect to use of training data, choice of algorithm, application of the algorithm to an individual's data, and use of the outcome of the algorithm in a decision.

Third, we recommend the use of counterfactual analysis in explaining algorithmic decisions. Models developed in this framework can be used to guide compliance with requirements arising from regulations such as the GDPR or governance by institutional review boards and other ethics review bodies. The GDPR requires data controllers to provide data subjects with "meaningful information about the logic involved" in an automated decision-making process, as well as information about "the significance and the envisaged consequences of such processing for the data subject" [9]. Providing data subjects with information regarding the logic involved in an algorithmic decision-making process can be informed by an analysis designed to help an individual understand why a particular decision was reached [27,28]. Informing data subjects about the envisaged consequences can be done in the form of a counterfactual analysis as described in this article, with assessment focusing on counterfactual causal estimation of the effects of an automated decision on the well-being of an individual. This analysis framework also can be used to characterize whether a decision can be expected to produce significant, long-term effects on an individual's life course, measured in terms of lifetime wealth, health, longevity, subjective life satisfaction, and ability to make ethically relevant choices. It can also be used to help identify what should

be considered a "decision" for the purposes of providing such an explanation, with the list of counterfactual categories provided in Section III as a guide. Further, this approach may be useful wherever a meaningful explanation of the consequences of an algorithmic decision is required—for example, to explain credit decisions under the Fair Credit Reporting Act [26], or to inform subjects of the consequences resulting from the use of their personal data as part of a consent process.

The proposed systematic approach to analyzing harm has numerous benefits. For instance, when applied ex-ante, it can help level the playing field for the process, imposing less burden on individuals to identify harm, giving groups and regulators more opportunity to identify issues, and enabling externalities and those indirectly affected to be taken into account. When used ex-post, the process provides a basis for auditing an automated decision. Although algorithms can lead to unforeseen consequences, serious attention to harm in advance, and the implementation of harm monitoring and reduction reveals an intent not to discriminate. It can also be used to help avoid the pitfalls of discriminatory effects and to calibrate socially appropriate trade-offs between accuracy and error.

References

[1] C. Muñoz, M. Smith, and D. J. Patil, *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*, 2016.

[2] P. Dixon and R. Gellman, *The Scoring of America: How Secret Consumer Scores Threaten Your Privacy and Your Future*, 2014.

[3] A. Campolo et al., *AI Now 2017 Report* (2017), https://ainowinstitute.org/AI_Now_2017_Report.pdf.

[4] C. Sandvig et al., "When the Algorithm Itself is a Racist: Diagnosing Ethical Harm in the Basic Components of Software," *International Journal of Communication*, vol. 10, 2016, pp. 4972-4990.

[5] S. Barocas and A. D. Selbst, "Big Data's Disparate Impact," *California Law Review*, vol. 104, 2016, pp. 671-732.

[6] D. J. Solove and D. Citron, "Risk and Anxiety: A Theory of Data Breach Harms," *Texas Law Review*, vol. 96, 2018, pp. 737-786.

[7] Office of the Secretary of Health, Education, and Welfare, *Ethical Principles and Guidelines for the Protection of Human Subjects of Research* (The Belmont Report), April 18, 1979.

[8] D. Dittrich and E. Kenneally, *The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research*, U.S. Department of Homeland Security Technical Report, Aug. 2012.

[9] General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679) art. 13(f), 14(g), 15(1)(h).

[10] J. Angwin et al., "Machine Bias," *ProPublica*, May 23, 2016.

[11] S. Wakefield, Criminal Justice and the Life Course, in *Handbook of the Life Course*, Vol. II (M. J. Shanahan, J. T. Mortimer, and M. Kirkpatrick Johnson, eds.), 2016.

[12] D. E. Roberts, "The Social and Moral Cost of Mass Incarceration in African American Communities," *Stanford Law Review*, vol. 56, 2014, pp. 1271-1305.

[13] A. Plagnol, "Subjective Well-being over the Life Course: Conceptualizations and Evaluations," *Social Research*, vol. 77, no. 2, 2010, pp. 749-768.

[14] R. Layard et al., "What Predicts a Successful Life? A Life-Course Model of Well-Being," *The Economic Journal*, vol. 124, Nov. 2014, pp. F720-738.

[15] A. Sen, *Inequality Reexamined*, Cambridge, MA: Harvard University Press, 1992.

[16] A. Benforado, *Unfair: The New Science of Criminal Justice*, New York, NY: Penguin Random House, 2015.

[17] D. B. Rubin, "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions," *Journal of the American Statistical Association*, vol. 100, no. 469, 2005, pp. 322-331.

[18] G. W. Imbens and D. B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge, UK: Cambridge University Press, 2015.
[19] E. Bareinboim and J. Pearl, "Causal inference and the data-fusion problem," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 27, 2016, pp. 7345-7352.
[20] C. Dwork et al., "Fairness Through Awareness," *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012.
[21] C. Dwork and D. K. Mulligan, "It's Not Privacy, and It's Not Fair," *Stanford Law Review Online*, vol. 66, 2013, pp. 35-40.
[22] *See* A. Acquisti, C. Taylor, and L. Wagman, "The Economics of Privacy," *Journal of Economic Literature*, vol. 52, no. 2, 2016, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2580411.
[23] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *Proceedings of Innovations in Theoretical Computer Science (ITCS)*, 2017.
[24] D. Pager, "The mark of a criminal record," *American Journal of Sociology*, vol. 108, 2003, pp. 937–975.
[25] J. A. Kroll et al., "Accountable Algorithms," *University of Pennsylvania Law Review,* vol. 165, 2017, pp. 633-705.
[26] Fair Credit Reporting Act, 15 U.S.C. § 1681.
[27] F. Doshi-Velez et al., "Accountability of AI under the Law: The Role of Explanation," Working paper, 2017; https://arxiv.org/abs/1711.01134.
[28] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR," *Harvard Journal of Law & Technology*, vol. 31, 2018, pp. 841-887, https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sarah-Wachter-et-al.pdf.