AI and Ethics
https://doi.org/10.1007/s43681-024-00473-y

ORIGINAL RESEARCH

Algorithmic evidence in U.S criminal sentencing

Suzanne Kawamleh[1]

Received: 22 January 2024 / Accepted: 19 March 2024

Abstract
The use of automated risk assessment tools to predict a defendant's risk of recidivism is necessarily unfair. There is a tradeoff between equal treatment and equal outcomes which constitutes the "impossibility of fairness" problem in machine learning. This article provides an account of algorithmic fairness that centers on equal treatment and requires the use of *equally confirmatory* algorithmic evidence. The analysis relies on a Bayesian account of evidence to assess AI predictions of recidivism risk as evidence for or against hypotheses about a black and white defendant's probability of future rearrest. Such predictions are shown to provide weaker confirmatory evidence for a black defendant's future recidivism risk than a white defendant. Thus, the use of such evidence is necessarily unfair to black defendants because such use violates equal treatment and thus cannot meet a necessary condition of algorithmic fairness. This proposed account of algorithmic fairness provides the theoretical resources to avoid the "impossibility of fairness" problem. On this view of algorithmic fairness, fairness is neither inevitable nor impossible. By requiring equally confirmatory scores, rather than simply the same scores, decision makers can both satisfy equal treatment and reduce racial disparities in criminal sentencing.

Keywords Artificial intelligence · Fairness · Criminal justice · Confirmation · Evidence

Abbreviations

AI	Artificial intelligence
ML	Machine learning
RAT	Risk assessment tool
COMPAS	Correctional offender management profiling for alternative sanctions
FML	Fair machine learning

1 Introduction

Artificial intelligence (AI), including machine learning (ML) systems, are now used for automated decision making and decision support at every step of the criminal justice system—from predicting an individual's propensity for future criminal activity to making parole recommendations [4, 9, 14, 16]. Consider the following examples. Predictive policing tools use ML to construct heat maps identifying the most likely areas of future criminal activities. Predictive assessment tools are used to flag individuals as being potential threats to national security and place them on No-Fly Lists. Pretrial risk assessment tools, which include information about race, gender, and socioeconomic status, but not the type of crime committed, are used to predict the probability of an individual showing up to trial and to justify pre-trial detention. Risk Assessment Tools (RATs) are used to predict an individual's probability of committing future crime and to assign them a risk score; that score, aggregated with other types of evidence, is used to make recommendations for setting bail, sentence type and length, and eligibility for parole and release. One widely used and controversial algorithm is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm. In short, the use of complex and opaque [24] algorithms permeates the U.S. criminal justice system.

The use of algorithmic systems for criminal sentencing has rightfully elicited a great deal of public criticism and concern. One especially influential and powerful line of criticism is that such algorithms are trained on data produced by racist institutions, encode value laden and racialized assumptions about crime and criminality, produce racially biased risk scores and predictions, and consequently result in the disparate treatment of black and white defendants, which perpetuates—as well as magnifies—structural and systematic injustice in the U.S. criminal justice system. Thus, most

✉ Suzanne Kawamleh
skawamle@iu.edu; suzannekawamleh@gmail.com

[1] Department of Philosophy, Indiana University, Bloomington, IN 47405, USA

Published online: 22 April 2024

of the legal and moral debate around the adoption and use of such systems centers on concerns about algorithmic bias and fairness [12, 22, 23, 29]. Such concerns around algorithmic bias reflect a deeply felt need to identify stronger conditions of fair use for such RATs and related algorithmic fairness conditions.

Given that these systems reflect, magnify, and propagate existing societal biases, including racial bias, the domain of fair machine learning (FML) and algorithmic fairness has focused on developing and evaluating technical methods for detecting bias, debiasing systems, and imposing the correct accuracy and/or fairness constraint on algorithmic predictions. The aim of this research program is to identify a mathematical constraint which, if satisfied, guarantees the fairness of an algorithmic system's prediction(s) according to some specified condition or fairness measure. However, multiple authors have proven the impossibility of satisfying multiple normatively desirable fairness conditions, which is known as the problem of impossible fairness [6, 18]. This has led to a stalemate in discussions of algorithmic fairness and fair machine learning, followed by calls for the development of more substantive accounts of algorithmic fairness [13].

This article provides an account of algorithmic fairness as the fair use of algorithmic scores as evidence. Rather than argue for or against a particular fairness measure or debate what the criteria and scope of algorithmic fairness ought to be, algorithmic outputs—including predictions—are recast as evidence. A Bayesian account of evidence is used to assess COMPAS predictions as evidence for or against conditional probabilities about a black and white defendant's probability of future rearrest given the COMPAS risk score. This analysis shows that COMPAS scores are strongly confirmatory for two conditional probabilities concerning recidivism risk: (1) A white defendant is more likely to reoffend given a COMPAS score of "high risk" and (2) A black defendant is less likely to reoffend given a COMPAS score of "high risk". However, COMPAS predictions are weakly confirmatory for two other conditional probabilities concerning recidivism risk: (3) A white defendant is less likely to reoffend given a COMPAS score of "low risk" and (4) A black defendant is more likely to reoffend given a COMPAS score of "high risk".

This confirmational analysis of COMPAS predictions has important ethical implications. On the following account of algorithmic fairness, the notion of equal treatment is explicated as the adoption of evidential standards that require equally confirmatory evidence for decisions that punish black and white defendants. Moreover, such standards need to be responsive to the practical stakes involved in a given evidence-based decision. Given the high stakes involved and the significant consequences of error, only equally *strong* confirmatory evidence should be used for pre-trial and sentencing decisions. It is fair to use strongly confirmatory scores, like COMPAS "low risk" scores as a mitigating factor in pre-sentencing and sentencing decisions for black defendants, and it is unfair to use weakly confirmatory scores, like COMPAS "high risk" scores, as an aggravating factor in pre-sentencing and sentencing decisions for black defendants.

Thus, on this view, fair data-driven decision-making centers on the fair use of algorithmic evidence for human decision-making. While the scope of this analysis is limited to the use of COMPAS risk scores for two racial groups, black and white defendants, the underlying principle regarding the equitable treatment of diverse individuals through the requirement of equally confirmatory evidence applies more generally.

2 Legal challenges to RATs and evidence-based sentencing (EBS)

COMPAS is a widely used risk assessment algorithm for predicting recidivism risk. COMPAS classifies defendants as being low risk, medium risk, or high risk for recidivism where recidivism is measured by probability of rearrest within the following two years. Such RATs are widely adopted as part of a growing trend towards Evidence-Based Sentencing:

> "'Evidence,' in this formulation, refers not to the evidence in the particular case but to empirical research on factors predicting criminal recidivism. Based on that research, EBS provides sentencing judges with risk scores for each defendant based on variables that, in addition to criminal history, often include gender, age, marital status, and socioeconomic factors such as employment and education. EBS has been widely hailed by judges, advocates, and scholars as representing hope for a new age of scientifically guided sentencing" [27], 85)

This risk score is then included in the defendant's presentence investigation report which is then used to guide a judge's sentencing decision [27], 809). For example, in 2013, Eric Loomis was charged with five criminal counts in connection with a drive-by shooting. Loomis accepted a plea deal, pleading guilty to lesser charges of attempting to flee a traffic officer and operating a motor vehicle without the owner's consent. In doing so, the prosecutor and defense attorney had agreed to a deal of one year in county jail with probation. However, before sentencing, a Wisconsin Department of Corrections officer introduced a presentence investigation report that included a risk assessment score computed by COMPAS. COMPAS identified Loomis as high risk for violence, recidivism, and pretrial flight risk. As a result, the

court then sentenced Loomis to six years in prison with five years of extended supervision. The court denied Loomis probation and explicitly mentioned COMPAS stating:

> "You're identified, through the COMPAS assessment, as an individual who is at high risk to the community. In terms of weighing the various factors, I'm ruling out probation because of the seriousness of the crime and because your history, your history on supervision, and the risk assessment tools that have been utilized, suggest that you're extremely high risk to re-offend" [12], 90).

This case in particular has elicited widespread legal debate concerning the constitutionality of RATs given a defendant's right to due process and equal protection [12, 20, 25]. Freeman [12] argues that the Court's argument and decision in State v. Loomis was incorrect:

> "the Wisconsin Supreme Court incorrectly assessed the impact of the COMPAS algorithm and that courts should not use risk assessment algorithms during the sentencing process without stronger due process protections in place, if courts are to use the algorithms at all…" [12], 78).

Importantly, the case decision did not impose adequate safeguards for the use of RATs in sentencing:

> "While the court had the opportunity to take a progressive step towards protecting defendants' due process rights, they instead incorrectly applied precedent and issued flimsy warning labels that offer little to no protection against the use of COMPAS" [12], 88).

Such warning labels are inadequate given the concerns with due process and the defendant's inability and lack of means to investigate any misinformation:

> "Without access to the source code of the algorithm, neither Loomis nor any other defendant truly has the "means" to investigate any potential misinformation. Since neither the court nor the defendants are certain of what goes into the calculation of risk scores, defendants can only present a superficial argument against the elements that may or may not be included in the algorithm…its solution [warning labels] does not address the fact that defendants do not have the resources at their disposal to effectively investigate the accuracy of COMPAS to the extent that due process requires" [12], 94-95).

Furthermore, legal scholars argue that the use of demographic, socioeconomic, neighborhood, and family variables to determine whether and how long a defendant is incarcerated violates a defendant's right to an individualized sentencing process [12], 96) and has led critics to characterize EBS as "the scientific rationalization of discrimination" [27]. Starr argues that reliance on RATs that use gender and socioeconomic variables "formally incorporates discrimination based on socioeconomic status and demographic categories into sentencing" [27], 821). which rely on statistical generalizations about groups (e.g., "men commit more violent crimes than women") and make predictions about average recidivism for a group of offenders sharing the defendant's characteristics (e.g., average recidivism rate for male offenders). The use of such generalizations as the basis of an individual prediction and sentence is constitutionally problematic because individuals have a right to equal and individual treatment. However, tools that rely on generalizations about group tendencies to crime for an individual risk score effectively punish an individual based on their group membership rather than their individual actions and characteristics: "Incarceration, after all, profoundly interferes with virtually every right the Supreme Court has deemed fundamental, and EBS makes these rights interferences turn on identity rather than criminal conduct" [27], 823).

This is especially problematic when such tools rely on static demographic features like race, age or gender, "variables that are outside the defendant's control, unchangeable, and often the basis for considerable social stigma and disadvantage" [27], 822). Starr considers the case of a human judge tasked with determining a sentence for an individual man. It would ordinarily be considered reprehensible—and discriminatory— for the judge to assign a man a more severe sentence than a woman who commits the same crime based on the fact that the offender is male, and, statistically, men are more likely to reoffend. Starr notes that while such implicit bias certainly occurs,

> "it is virtually unheard of for modern judges to say they are taking gender into account…Until the past few decades, explicit consideration of gender as well as race was common, but few today defend that practice. The Federal Sentencing Guidelines, for example, expressly forbid the consideration of both race and sex. Outside the literature on EBS, scholars have likewise mostly treated the gender gap as an 'unwarranted' sentencing disparity" [27], 825).

Yet, discrimination based on race and gender is widespread in RATs which routinely classify black and/or male defendants as higher risk than white men and/or female defendants. Starr persuasively shows that "sentencing based on such instruments amounts to overt discrimination based on demographics and socioeconomic status" and is unconstitutional given individual's right to equal protection under the law and to be treated as individuals rather than merely as members of groups, and the average criminal tendencies thereof. This leads to a central legal and ethical criticism of

COMPAS, namely that its use perpetuates racial bias against black defendants and is thus unfair [2, 12, 23].

3 Algorithmic (racial) bias and fairness

An independent analysis by ProPublica [1] revealed that COMPAS was racially biased against black defendants. COMPAS misclassified black defendants as high risk at twice the rate of white defendants, and misclassified white defendants as low risk at twice the rate of black defendants. Defenders of COMPAS do not dispute this finding, but rather argue that it is unreasonable to expect COMPAS to display the same sensitivity and specificity for black and white defendants, given different base rates of recidivism among black and white defendants [10]. This is important because these errors have very different consequences. False positive error rates can lead to harsher and longer sentences, denial of parole or bail, etc. False negative error rates can lead to shorter or more lenient sentences, bail, early parole, etc. The disparity in error rates of predictions increase disparities in the distribution of punishments and allowances for black and white defendants who are accused of committing the same criminal offense. In this way, it reinforces and magnifies the existing racial disparities in criminal sentencing and long-standing systematic biases in criminal justice processes and institutions against black citizens.

ProPublica analysts demonstrated that COMPAS failed to satisfy error-rate balance, a statistical measure of fairness that seeks to produce equal outcomes for black and white defendants. Given the racial basis for misclassification and the asymmetrical negative consequences and harms associated with being misclassified as high risk (as opposed to low risk), ProPublica analysts argued that COMPAS was biased and unfair to black defendants.

NorthPointe Corporation, the developers of COMPAS, pushed back and argued that, in fact, COMPAS was fair because it satisfied a different statistical measure of algorithmic fairness, predictive parity, which ensures the equal treatment of black and white defendants [10]. The COMPAS risk score is effectively independent of race. A score of 0.6 for a black defendant *means the same thing* as a risk score of 0.6 for a white defendant; namely, it means a defendant has a 60 percent chance of recidivism, whether the defendant is black or white. Furthermore, if the algorithm satisfies the equal treatment of black and white defendants, as COMPAS does, it will necessarily produce unequal outcomes for black and white defendants. However, NorthPointe argues that such unequal outcomes are not unfair but a necessary result of the fair treatment of individual defendants drawn from populations with different base rates of recidivism. In other words, they appeal to the "impossibility of fairness" to justify the fairness of COMPAS for predicting recidivism risk, despite the mis- and over-classification of black defendants as high risk.

Independent scholars have argued that the dispute boils down to a disagreement about the right definition of fairness and that the fairness of COMPAS was subject to reasonable disagreement, as neither definition of fairness was clearly nor objectively correct [7, 11, 18]. Unequal outcomes are the unintentional, but also unavoidable, cost of a fundamental commitment to equal treatment and the more basic mathematical tension between satisfying conditions of equal treatment and equal outcomes. To secure equal outcomes, and thus satisfy fairness as defined by the ProPublica critics, the algorithm would have to violate predictive parity and assign risk scores based, in part, on race. This would be discriminatory. Alternatively, to uphold equal treatment, and thus satisfy fairness as defined by NorthPointe Corporation, the algorithm would have to assign risk scores to defendant profiles independently of race. However, treating black and white defendants equally will necessarily produce unequal outcomes in the form of imbalanced error rates. The fact that false positive errors are more likely for black defendants is not unfair but merely reflects the unequal base rate of recidivism in black and white populations where recidivism rates are higher in black populations. Or so defendants of COMPAS have argued.

Given this dispute over the appropriate measures of algorithmic fairness and the impossibility of satisfying any two measures of fairness simultaneously, some have concluded that "total fairness cannot be achieved" [3] and "it is highly unlikely that an algorithmic justice approach will advance" [8].

One reason why the dispute about algorithmic fairness has reached such an impasse is that fairness has been construed as technical fairness—fairness as a mathematical property of an algorithmic model. The question of fairness is reduced to a question of which mathematical constraints should be imposed on algorithmic prediction to guarantee a "fair" prediction. This construal of the problem of algorithmic fairness as a technical problem is misguided. There is no way to design an algorithm to ensure a fair outcome. This is because fairness and justice are not reducible to statistical parity or other formal mathematical constraints. There is no formula or algorithm that can calculate what a fair decision would be just as there is no mathematical or algorithmic solution to the moral problems of unfairness, racism, or social injustice.

Moreover, there is a conceptual confusion at the heart of debates about algorithmic fairness: accuracy is a property of algorithmic predictions while fairness is a property of human decisions made using or based on algorithmic predictions. Much of the debate about algorithmic fairness and fair machine learning methods concern incompatible measures of algorithmic accuracy. However, the incompatibility

of different accuracy measures does not entail the "impossibility of fairness" in decision making. Accurate predictions are neither necessary nor sufficient for fair decision-making. Consider the following. Even if an algorithm could be developed to satisfy both conditions of predictive parity and error rate balance, this would not entail that decisions made based on those algorithmic predictions are fair or just. Accurate genomic predictions can still be used for making unfair and unjust decisions to sterilize individuals against their will for eugenicist and racist purposes. The accuracy of such predictions would not mitigate, but rather exacerbate, the problem of racism or social injustice. Conversely, one can make fair and just decisions without very accurate predictions. For example, a community can collectively vote to adopt climate adaptation measures that are robust to a host of different possible climate conditions and such a decision can be fair without requiring very precise or accurate predictions of future precipitation or heat levels. It is wholly consistent for an individual to hold that an algorithm makes an accurate prediction, on a given measure of accuracy, but that it is nonetheless unethical to use that prediction as a basis for making certain decisions and vice versa. The point is that there is an important gap between prediction and decision. Improving the accuracy or "fairness" of predictions does not necessarily improve the fairness or effectiveness of decisions, even when such decisions are made based on the predictions. An account of algorithmic fairness should bridge this gap and mediate the relation between algorithmic prediction and algorithm-based decision making.

This distinction between accurate (or "fair") predictions and fair prediction-based decisions is important because it allows us to advance the discourse concerning algorithmic fairness and justice beyond the current stalemate of "impossible fairness" in computer science. The technical debate about the merits of predictive parity and equalized odds (*i.e.*, error-rate balance) conditions will be set aside for the time being, with the focus being on whether recidivism risk scores provide good evidence for judicial decision making.

4 Assessing COMPAS predictions as evidence for decision-making

4.1 COMPAS risk scores as evidence

COMPAS classifies defendants as being low risk or high risk for recidivism where recidivism is measured by probability of rearrest in the following two years. This risk classification represents a conditional probability that the defendant will reoffend given certain input features to COMPAS. In other words, the COMPAS score provides evidence (e) about the conditional probability that a defendant will reoffend and be rearrested (H|e). Given the fact that the algorithmic output concerns risk, it can intuitively be recast as a prediction. The algorithm makes a prediction about the probability of an individual defendant being rearrested within two years. The prediction itself is often used as evidence by a judge for making pre-trial and sentencing decisions. The judge effectively marshals the algorithmic prediction as evidence, alongside other forms of evidence, for or against a pre-sentencing or sentencing judgment. Thus, the algorithm is also an evidence-generating process where risk predictions are often used as legal evidence to justify adopting a harsher sentence or lighter sentence, subject a defendant to pre-trial detention or allow release pending bail, to justify early release decisions, etc.

By reconceptualizing risk scores as evidence, for both a conditional probability and sentencing decision, one can identify the conditions under which the use of this evidence is fair and connect the dispute over fairness in machine learning to a rich existing literature on procedural fairness, legal evidence, ethical standards of evidential justification, etc. Furthermore, the concept of evidence is more appropriate for understanding the mediating role of prediction in algorithm-based decision-making. Instead of asking whether a prediction is "fair" in some formal and acontextual sense, this question can be recast in a clearer and better-defined way: does the prediction provide good evidence for a given decision? This leaves us with a question of how experts should assess the adequacy of algorithmic evidence, such as COMPAS risk scores, for decision making purposes.

4.2 Confirmation theory

Probabilistic and statistical methods, of which machine learning is a branch, function as tools that produce evidence for empirical hypotheses, e.g., a defendant can be safely released. The degree and type of justification provided by such evidence has been a long-standing subject of philosophical interest. The Bayesian approach to confirmation theory is both widely accepted and well-established as the dominant approach among philosophers of science.

Confirmation theorists define and analyze evidence based on the degree to which evidence (e) increases (*i.e.*, confirms) or decreases (*i.e.*, disconfirms) the probability (p) of a given hypothesis (H). Thus, on this view, if e is confirmatory evidence for H, then the conditional probability $p(H/e) > p(H)$ and if e is dis-confirmatory evidence for H, then $p(H) > p(H/e)$. An algorithmic classification or prediction can function as evidence for a conditional probability. Where the algorithmic prediction provides confirmatory evidence, the defendant is more likely to reoffend given a prediction of "high risk" than otherwise, so that the $p(H|e) > p(H)$. Where the algorithmic prediction provides disconfirmatory evidence, the defendant is less likely to be rearrested given the prediction of "high risk" than otherwise,

so that p(H|e) < p(H). The confirmation of such conditional probabilities is then taken to be evidence for a further hypothesis (H) that, for example, "defendant X poses a public risk and should be incarcerated pending trial".

4.3 Relative confirmatory strength of COMPAS evidence

Let us begin by reviewing the performance of a widely used recidivism prediction algorithm, COMPAS. Larson et al. [19] tested the accuracy of COMPAS for 11,757 individuals arrested and assigned a COMPAS score in Broward County, FL between 2013 and 2014. COMPAS developers define recidivism as a finger-printable arrest that results in a jail booking within two years after the COMPAS score. Larson et al. [19] compared these 11,757 individuals' scores with subsequent public criminal records of each individual to check whether their records of incarceration confirmed the COMPAS score they received and related predictions about recidivism in the form of rearrest. They then compared the COMPAS accuracy and error rates for black and white defendants (See [19] for details of analysis). They found that: "In forecasting who would re-offend, the algorithm correctly predicted recidivism for black and white defendants at roughly the same rate (59 percent for white defendants, and 63 percent for black defendants)." [19]. They also found that COMPAS predictions fail differently for black and white defendants. COMPAS incorrectly labeled black defendants as high risk at a rate of 45 percent compared to 23 percent of white defendants. In other words, the false positive error rate for black defendants was almost twice that of white defendants. On the other hand, COMPAS incorrectly labeled white defendants as low risk at a rate of 48 percent compared to 28 percent of black defendants. In other words, the false negative error rate for white defendants was almost twice that of black defendants. Finally, they found that even when they controlled for age, gender, prior crimes, etc., black defendants were 45% more likely to be assigned a higher risk score than a white defendant and 77% more likely to be assigned a higher risk score for violent recidivism than a white defendant. In probabilistic terms, p(e), where e is a high risk score, is consistently, and significantly, higher for black defendants than white defendants. This has important implications for the use of the evidence to discriminate between alternative hypotheses (H and ~H) for black defendants.

These results can be summarized as the true and false positive and true and false negative rates of COMPAS predictions of recidivism. The true positive rate (TPR) is the ratio of individuals who COMPAS correctly classified as high risk, or true positives (TP), to the total number of individuals who actually recidivated, both true positives and false negatives (FN). This is represented as TPR = [TP / (TP + FN)]. The false positive rate (FPR) is the ratio of individuals who COMPAS wrongly classified as high risk, or false positives (FP), to the total number of individuals who actually did not recidivate, both false positives and true negatives (TN). This is represented as FPR = [FP/(FP + TN)]. The true negative rate (TNR) is the ratio of individuals who COMPAS correctly classified as low risk, or true negatives, to the total number of individuals who actually did not recidivate, both true negatives and false positives. This is represented as TNR = [TN/(TN + FP)]. Finally, the false negative rate (FNR) is the ratio of individuals who COMPAS misclassified as low risk, or false negatives (FN), to the total number of individuals who actually recidivated, both true positives and false negatives. This is represented as FNR = [FN/(FN + TP)].

On the Bayesian view of confirmation, the COMPAS risk score is strong evidence if it significantly increases the conditional probability that a defendant will be rearrested in the next two years given the algorithmic risk prediction of "high risk". Evidence is confirmatory if p(H|e)» p(H). Given Bayes' theorem, this can also be restated as p(e|H)» p(e|-H) where p(e|-H) reflects the probability of the evidence given the null hypothesis that a defendant is not rearrested. To make the Bayesian argument that COMPAS risk scores confirm, and thus warrant increased confidence in, the prediction that a defendant will be rearrested in the next two years, the following premises must be true, where e = COMPAS score of "high risk":

(1) if p(H|e) » p(H), then e strongly confirms and warrants increased confidence in prediction H (defendant will recidivate)
(2) p(H|e) » p(H); given a COMPAS score of "high risk", a defendant is more likely to recidivate than otherwise.

Therefore, e strongly confirms and warrants increased confidence in H.

This argument is clearly valid, the question is whether the premises are in fact true for COMPAS and similar recidivism risk assessment algorithms for predictive hypotheses about a defendant's future rearrest. The first premise restates a central thesis of Bayesian inference and seems straightforwardly unproblematic. It is the second premise which needs to be defended. Is it in fact the case that a COMPAS risk score of "high risk" confirms or makes it more probable that a defendant will be rearrested in the future than otherwise?

Recall that the true positive rate for COMPAS risk scores was reported as 59% for white defendants and 63% for black defendants. This means that the ratio of correct predictions of recidivism to total actual recidivism was 0.59 among white defendants and 0.63 among black defendants. Importantly, this number drops dramatically for predictions of violent recidivism—COMPAS correctly predicts violent recidivism only 20% of the time or at a

rate of 0.2. Importantly, these rates fall short of the evidential standard that Northpointe developers themselves adopt: "A rule of thumb according to several recent articles is that AUCs of 0.70 or above typically indicate satisfactory predictive accuracy, and measures between 0.60 and 0.70 suggest low to moderate predictive accuracy" (Larson, et. al., 2016). However, the above rates ranging between 59 and 63% are well below 70%. This is important because the evidence in question is confirmatory to some degree (59–63%) but is not significantly confirmatory or confirmatory enough (greater than or equal to 70%), even on the Northpointe developers' standards. Thus, such evidence cannot warrant punitive decisions. Throughout the remainder of this analysis, the stated evidential standard of 70% will be adopted.

This is especially so given the realities of over-policing in predominantly black neighborhoods, the higher probability of arrest for black individuals than white individuals for the same actions, and the criminogenic effect of risk classifications and incarceration which disproportionately affect black Americans. Given these different sociopolitical conditions which directly affect rates of rearrest, it is unclear to what degree the evidence from the COMPAS prediction actually makes it more likely that a black defendant will be rearrested because it is tracking some truth about criminal tendencies or recidivism as opposed to being the result of a confounding factors like racist surveillance, and policing and arrest patterns. While COMPAS predictions are usually taken to be evidence of the defendant's criminality and future criminal behavior, it may in fact provide evidence for the systematically racist tendencies of law enforcement and the role of the criminal justice system as a structural causal factor of rearrest rather than the criminality or danger posed by the defendant [13].

The natural follow-up question is whether the evidence is equally confirmatory for black and white defendants. Given the race-specific variations in COMPAS performance, experts need to evaluate this premise with respect to two predictive hypotheses which respectively specify a black and white defendant:

H_1 = "White defendant W_1 will be rearrested in the next two years".

H_2 = "Black defendant B_1 will be rearrested in the next two years".

Put in probabilistic terms, they need to compare $p(H_1|e)$ and $p(H_2|e)$ relative to $p(H_1)$ and $p(H_2)$. The sensitivity and specificity of COMPAS risk scores can be provided in the form of the true positive rate and false negative rate, respectively, and the baseline rate of recidivism and non-recidivism is provided. Therefore, the posterior probability of a defendant recidivating given the evidence provided by a COMPAS risk score can be calculated using the following form of Bayes' theorem:

In Fig. 1, p(e|H) refers to the true positive rate, p(e|~H) refers to the false positive rate, and p(H) and p(~H) refer to baseline rates of recidivism and non-recidivism in a sample.

In keeping with Larson et al.'s [19] methodology, all probability calculations will be performed relative to the subpopulation rather than the total population. For example, in the contingency table below, Larson et al. calculate the false positive rate of black defendants using a ratio of the number of black defendants who receive a high score and did not recidivate, or survived, (805) to the total number of black defendants who did not recidivate (805 + 990 = 1795) so the FPR = 805/1795 = 0.44846 = 44.85%. Similarly, the p(H), p(e|H), p(e), and p(H|e) will be calculated relative to the subpopulation. This will be reflected in the probability notation that uses subscripts to specify subgroup. For example, the prior probability of recidivism is calculated relative to the subgroup, so $p(H_1)$ is used for white defendants and $p(H_2)$ is used for black defendants, and likewise for other probabilities.

These probabilities will be used to calculate the posterior probability p(H|E) that a defendant recidivates given the COMPAS risk score for both black and white defendants. This will indicate whether the evidence of a COMPAS score confirms, or increases the probability of, a hypothesis about a defendant's recidivism. If p(H|e)» p(H), then the evidence confirms, and thus warrants increased confidence in, a predictive hypothesis that a defendant will recidivate. Then, it is necessary to determine how much the evidence confirms, or increases the probability of, a hypothesis or prediction H by calculating the difference between the posterior probability p(H|e) and prior probability (H). This will be referred to as the "confirmational significance" of the COMPAS evidence. Finally, there will be a comparison of the confirmational significance of high and low COMPAS risk scores for predicting the future recidivism of black and white defendants.

Let us consider the following. Figure 2 shows the contingency tables used by Larson et al. [19] in their analysis of COMPAS score performance:

The total number of defendants in the sample analyzed is 7,214, with 3,696 black defendants and 2,454 white defendants. The baseline recidivism rate of black defendants in this sample provides the prior probability of $p(H_2)$ which is (532 + 1369)/3696 = 0.5143 = 51.43%. The probability of a high-risk score, conditioned on a black defendant actually recidivating, is $p(e_2|H_2)$ is equivalent to the true positive rate which is the ratio of true positives to actual positives. In this case, TPR = 1369/(1369 + 532) = 1369/1901 = 0.72 or 72%. The baseline non-recidivism rate, or $p(\sim H_2)$, is (990 + 805)/3696

$$p(H|e) = \frac{p(H) * p(e|H)}{[p(H) * p(e|H) + p(\sim H) * p(e|\sim H)]}$$

Fig. 1 Bayes' theorem

All Defendants	Low	High
Survived	2681	1282
Recidivated	1216	2035
FP rate: 32.35		
FN rate: 37.40		
PPV: 0.61		
NPV: 0.69		
LR+: 1.94		
LR-: 0.55		

Black Defendants	Low	High
Survived	990	805
Recidivated	532	1369
FP rate: 44.85		
FN rate: 27.99		
PPV: 0.63		
NPV: 0.65		
LR+: 1.61		
LR-: 0.51		

White Defendants	Low	High
Survived	1139	349
Recidivated	461	505
FP rate: 23.45		
FN rate: 47.72		
PPV: 0.59		
NPV: 0.71		
LR+: 2.23		
LR-: 0.62		

Fig. 2 Contingency tables representing COMPAS score performance [19]

=0.485=48.5%. The probability of a high risk score, conditioned on a defendant not recidivating, is $p(e_2|\sim H_2)$ which is equivalent to the false positive rate=805/(990+805)=805/1795=0.4485=44.85%.

The baseline recidivism rate of white defendants, or $p(H_1)$, in this sample is (461+505)/2454=0.3936=39.36%. The probability of a high-risk score, conditioned on a white defendant actually recidivating, is $p(e_1|H_1)$ is equivalent to the true positive rate which is the ratio of true positives to actual positives. In this case, TPR=505/(505+461)=505/966=0.52 or 52%. The baseline non-recidivism rate, or $p(\sim H_1)$, is (1139+349)/2454=1488/2454=0.606=60.6%. The probability of a high risk score, conditioned on a defendant not recidivating, is $p(e_1|\sim H_1)$ is equivalent to the false positive rate which is 349/(1139+349)=349/1488=0.2345=23.45%.

Now that the values for p(e), p(h), and p(e|H), for black and white defendants are found, one can calculate each groups' respective p(H|e) using Bayes' Theorem. For black defendants, $p(H_2|e_2)$=(0.5143*0.72)/[(0.5143*0.72)+(0.485*0.4485)]=0.367/(0.367+0.2175)=0.367/0.5845=0.6278=62.78%. For white defendants, $p(H_1|e_1)$=(0.3936*0.52)/[(0.3936*0.52)+(0.606*0.2345)]=0.204672 / (0.204672+0.1421)=0.204672/0.346779=0.59=59%. Now, to determine the confirmational significance of the COMPAS evidence, compare p(H) to p(H|e) for both black and white defendants. For black defendants, $p(H_2|e_2)$=0.6278 and $p(H_2)$=0.5143 so 0.6278—0.5143=0.1135=11.35%. For white defendants, $p(H_1|e_1)$=0.59 and $p(H_1)$=0.3936 so 0.59—0.3936=0.1964=19.64%.

A high risk COMPAS score increases the probability that both black (~11%) and white defendants (~20%) recidivate but by very different amounts. A high risk COMPAS score is almost twice as confirmatory of a prediction concerning a white defendant's recidivism than a black defendant. It is clear is that the evidence provided by COMPAS is much more confirmatory for predictive hypotheses about white defendants' probability of recidivism than black defendants. This means that given the probabilities involved, a COMPAS risk score of "high risk" provides much stronger evidence and warrants greater confidence that a white defendant will reoffend and be rearrested than the same evidence would provide with respect to a black defendant. This difference in confirmational significance should be weighted accordingly in expert judgments.

Now, let us consider the strength of evidence provided by a COMPAS prediction of "low risk" for a black and white defendant. In this case, e is the statement "COMPAS labels a defendant as 'low risk'". To establish that e confirms, and thus warrants increased confidence in, a predictive hypothesis H that a defendant will *not* be rearrested in the next two years, p(H|e)» p(H).

Again, the total number of defendants in the sample analyzed is 7,214, with 3,696 black defendants and 2454 white defendants. The baseline non-recidivism rate, or $p(H_2)$, of black defendants in this sample is 48.5%. The probability of a low-risk score, conditioned on a black defendant not recidivating $p(e_2|H_2)$ is the true negative rate=990/(990+805)=990/1795=0.5515=55.15%. The baseline recidivism rate of black defendants in this sample $p(\sim H_2)$ is (532+1369)/3696=0.5143=51.43%. The probability of a low-risk score, conditioned on a black defendant actually recidivating, is $p(e_2|\sim H_2)$ is equivalent to the false negative rate which is 532/(532+1369)=532/1901=0.2799=27.99%. For black defendants, $p(H_2|e_2)$=(0.485*0.5515)/[(0.485*0.5515)+(0.5143*0.2799)]=0.2675/ (0.2675+0.1440)=0.2675/0.4115=0.65=65%.

The baseline non-recidivism rate of white defendants, or $p(H_1)$, in this sample is 60.6%. The probability of a low-risk score, conditioned on a white defendant not recidivating, $p(e_1|H_1)$, is the true negative rate of 1139/(1139+349)=1139/1488=0.7655=76.55%. The baseline recidivism rate of white defendants, $p(\sim H_1)$, in this sample is 39.36%. The probability of a low risk score, conditioned on a defendant recidivating, $p(e_1|\sim H_1)$, is equivalent to the false negative rate which is 47.72%. For white defendants, the $p(H_1|e_1)$=(0.606*0.7655)/[(0.606*0.7655)+(0.3936*0.4772)]=0.4639/(0.4639+0.1879)=0.4639/0.6517=0.7118=71.18%.

For a black defendant, $p(H_2|e_2) = 0.65$ and $p(H_2) = 0.485$, the difference being $0.165 = 16.5\%$. For a white defendant, $p(H_1|e_1) = 0.7118$ and $p(H_1) = 0.606$ so the difference is $0.1058 = 10.58\%$. In other words, the evidence provided by a COMPAS risk score of "low risk" is 6% more confirmatory for black defendants than it is for white defendants.

Figure 3 summarizes the confirmational significance of the different types of evidence that COMPAS risk scores can provide for black and white defendants:

One can rank the COMPAS risk score evidence by degree of confirmational strength, from strongest evidence to weakest evidence as follows:

1. COMPAS score of "high risk" for white defendant.
2. COMPAS score of "low risk" for black defendant.
3. COMPAS score of "high risk" for black defendant.
4. COMPAS score of "low risk" for white defendant.

There are a few important consequences that follow from this analysis of COMPAS risk scores as evidence. First, the confirmatory strength of COMPAS risk scores is highly race-dependent. The same risk score can provide strongly confirmatory evidence for a predictive hypothesis concerning a black defendant but poor evidence for the same predictive hypothesis when it concerns a white defendant.

Second, the confirmatory value of COMPAS evidence for race-specific predictive hypotheses has important epistemic implications on the aggregation of COMPAS evidence and other legal evidence, the use of COMPAS evidence as a basis for decision making, and the determination of appropriate evidential standards for machine evidence in criminal justice.

Third, this evidential analysis of COMPAS predictions has important ethical implications for procedural fairness in judicial decision making. Decision-makers, including judges, have an epistemic and ethical duty to use reliable and good evidence when making high-stakes decisions with significant and lasting consequences. Using weak or unreliable evidence as a basis for significant pre-sentencing and sentencing decisions raises ethical concerns and may undermine the fairness of the legal process. Thus, judges ought to use COMPAS predictions only if they provide a strong source of evidence for pre-sentencing and sentencing decisions. Conversely, judges should not use COMPAS predictions that provide weak and unreliable evidence for a given hypothesis and such use is morally blameworthy. In short, the epistemic assessment of algorithmic predictions as evidence can provide ethical and actionable guidance on the appropriate and fair use of algorithmic systems and their predictions in criminal justice.

Each of these consequences will be elaborated upon in the next section.

5 Evidence and ethics in data-driven decision-making

5.1 Inductive risk and evidential standards

Given the serious punitive consequences associated with decisions made based on a "high risk" classification, and errors thereof, there ought to be a high, and higher, threshold for when a "high risk" classification is deemed good evidence (in comparison to a "low risk" classification).

It is widely accepted that epistemic standards should be responsive to the practical stakes at hand so that high stakes decisions require better, more reliable, evidence than inconsequential decisions. This can be fleshed out in terms of an expected value approach. One can assign a value to different expected outcomes, weight such outcomes by the probability of their occurrence, then sum the weighted values. Given that injustice has a high negative expected value, this indicates a need for a higher standard of evidence that the judgment in question is not unfair in order to make decisions with a higher expected value.

In philosophy of science, a similar argument is made with respect to inductive risk. Given the fallibility of any hypothesis, the decision to accept or reject a hypothesis depends on a determination of whether the evidence provided is strong enough to justify acceptance or rejection. This decision is underdetermined by the facts and depends in a strong way on a moral judgment about the significance of making a mistake in either accepting or rejecting a hypothesis for the purposes of decision making and action. In other words, such a judgment is fundamentally value-based and reflects a moral judgment about the acceptability of error, or risk of error, and the consequences thereof [26]. If the consequences of

Race	COMPAS Score	Confirmational Value
Black	Low Risk	16.5%
White	Low Risk	10.5%
Black	High Risk	11.3%
White	High Risk	19.6%

Fig. 3 Table summarizing the confirmational value of COMPAS risk score evidence for black and white defendants

error in hypothesis acceptance (or rejection) are significant, then there should be a higher evidential standard than if the consequences of error are less important or insignificant.

Given the significant and lasting consequences of arrest, detention, and incarceration— and wrongful arrest, detention and incarceration—the evidential standards for using COMPAS scores of "high risk" as an aggravating factor in pre-sentencing or criminal sentencing should require (1) strongly confirmatory evidence (PPV > 0.7) and (2) stronger confirmatory evidence than the evidential standards for the use of COMPAS scores of "low risk" as a mitigating factor to justify a reduced sentence, parole, or early release. What this means is that whatever evidential standard or threshold is adopted, the evidential standard for COMPAS scores of "high risk" should generally be higher than those of "low risk" scores because better evidence is required to impose harm on an individual than to grant them an allowance.

More importantly, the fair use of evidence requires the equal evidential treatment of black and white defendants. On this evidentialist account of algorithmic fairness, the notion of equal treatment is explicated as the adoption of evidential standards that require *equally confirmatory evidence* for decisions that punish black and white defendants. The threshold for evidence to be confirmatory enough for a punitive decision should be sufficiently high, as previously discussed (PPV ≥ 0.7), and equally high for black and white defendants. Assuming that to be the case, the evidence produced by COMPAS scores of "high risk" can meet such an evidential standard in the case of white defendants, but not in the case of black defendants.

Consider, for argument's sake, that the threshold for confirmational significance is set at 15 percent (0.15) for a COMPAS "low risk" score and at 20 percent (0.2) or 30 percent (0.3) for for a COMPAS "high risk" score. The confirmational significance of "low risk" score evidence for white defendants is 10.6%. This would mean that COMPAS "low risk" scores do not provide confirmatory evidence for the low risk of recidivism of white defendants and cannot warrant the use of COMPAS risk scores as a mitigating factor in sentencing white defendants. The confirmational value of "low risk" score evidence for black defendants is 16.5 percent. As such, a COMPAS "low risk" score does provide confirmatory evidence for the low risk of recidivism of black defendants and should be used as a mitigating factor in sentencing black defendants.

Given our hypothetical 20 percent or 30 percent cutoff for the confirmational significance of "high risk" score evidence, the confirmational significance of "high risk" score evidence for black defendants is too low at 11.3 percent. With respect to COMPAS scores of "high risk", such scores do not meet the necessary confirmational threshold for black [or any] defendants and cannot be ethically used as evidence against, or as an aggravating factor in pre-sentencing and sentencing, black defendants. For example, such a COMPAS designation of high risk should not be used to justify denying bail for a black defendant (and perhaps any defendant). However, based on the same threshold cutoff, such evidence is fairly confirmatory of hypotheses concerning white defendants. The confirmational value of "high risk" score evidence for white defendants is 19.6 percent, almost 20%, and may warrant the use of such scores as an aggravating factor in sentencing white defendants. Alternatively, if the threshold for confirmational significance of COMPAS evidence of high recidivism was 30 percent or greater, then COMPAS scores would provide poor evidence that is not confirmatory enough to warrant punitive decisions for either black or white defendants.

In short, the equal treatment of black and white defendants should mean equal evidential standards of confirmational strength. Algorithmic fairness, on this view, means requiring that the evidence provided by COMPAS risk scores meet equal standards of confirmational significance for black and white defendants. This may, as in this case, entail weighting the same evidence (e.g., COMPAS risk score) differently for members of different groups. The weight assigned to COMPAS evidence, like "high risk" classifications, should reflect the confirmational significance of the evidence where such significance is highly context- and race-sensitive. Given the greater confirmational significance of such evidence for white defendants compared to black defendants, COMPAS scores of "high risk" should be assigned greater weight when assessing the risk of white defendants. This is because the evidence of a "high risk" classification is more confirmatory of the hypothesized recidivism of white defendants than black defendants. Similarly, COMPAS evidence in the form of "low risk" classifications should be assigned greater weight when assessing the risk of black defendants. This is because the evidence of a "low risk" classification is more confirmatory of the hypothesized non-recidivism of black defendants than white defendants. Thus, setting equal thresholds for confirmational significance of COMPAS evidence will lead to the same evidence being weighted differently for members of different racial subgroups. By making equal confirmatory strength a defining feature of fairness in evidence use, this account of algorithmic fairness can both uphold the normatively desirable principle of equal treatment while being sensitive to ways in which the same evidence is more or less confirmatory for members of different racial groups.

This type of context-sensitivity is critical to the equitable, and thus fair, treatment of defendants from different racial groups. It reflects an important distinction between equality, equity, and fairness. Equality involves treating everyone the same regardless of the circumstances or their needs and equity refers to treating individuals based on their specific circumstances or needs. Fairness requires ensuring just or

appropriate treatment given the context and circumstances, thus calling for equitable treatment rather than equal treatment when the circumstances or context are very different. The use of equal evidential standards of confirmational strength rather than equal risk scores is critical to the equitable, and thus fair, use of COMPAS risk scores in decision-relevant contexts.

This also means that the equal weighting of COMPAS evidence within the criminal justice system is unfair when that evidence exhibits different confirmational strength. Such "equal treatment" ignores the unequal confirmational strength of algorithmic evidence and violates our notion of algorithmic fairness which demands equally confirmatory evidence for warranting judicial decision making and sentencing. It is both epistemically and morally inappropriate to use poor quality evidence, such as a COMPAS score of "high risk", for imposing a harsher sentence on a black defendant, or a COMPAS score of "low risk" for reducing the sentence of a white defendant. Moreover, given that stronger evidence is needed to warrant imposing a punishment than to provide an affordance or allowance in the form of a reduced sentence, the harm of using "high risk" scores to justify the pre-trial incarceration of black defendants is greater than the harm of using "low risk" scores to justify offering bail or reducing a sentence for a white defendant. This implies that the use of "high risk" scores as an aggravating factor for sentencing black defendants is open to both epistemic and ethical reproach.

The thresholds considered above may be too low. Perhaps much greater thresholds of confirmatory value are needed for what is legal evidence in criminal proceedings. In this case, perhaps the evidence provided by COMPAS is not strong enough in any of these cases, for either black or white defendants. However, this analysis draws attention to the population-relative dynamics of machine evidence in criminal justice. Moreover, this philosophical analysis provides the resources to point out ways in which such risk scores can be used, and abused, as legal evidence in decision making. Weak evidence should not be used to detain people, strongly confirmatory evidence is needed when the risks of harm are very high for defendants, and the evidence required to imprison people should be stronger than the evidence required to reduce a sentence.

Moreover, such an analysis avoids problems of unequal treatment with different race-specific thresholds of riskiness where the quantitative threshold for a black defendant to be considered high risk is higher than that of a white defendant, a proposal termed "affirmative algorithms" [14, 15]. Instead, decision makers can adopt equal treatment with respect to the confirmational strength of the evidence provided by COMPAS scores without race-norming the scoring algorithm or the scores directly. Nonetheless, by making predictive hypotheses sensitive to the racial identity of the defendant, such an approach is race-sensitive without violating equal-protection laws and enables decision makers to uphold a shared and common threshold for the confirmatory value of the evidence for both black and white defendants. For example, the threshold for confirmational significance of evidence of "low risk" can be the same for black and white defendants while the confirmational significance of the evidence itself is race-sensitive.

6 Aggregating machine and legal evidence

Treating COMPAS risk scores as evidence and analyzing the confirmatory value of such evidence also has important consequences for the activity of aggregating evidence, machine and otherwise, in judicial discretion and judgment. Pre-sentencing and sentencing decisions should not rely solely or substantially on algorithmic risk scores. Despite justices' attempts to limit the role of COMPAS risk scores as corroborative rather than determinative [21], in practice these scores often play an outsized role in judicial decision making. Consider the following case of a construction worker, Zilly, who was given the maximum sentence for stealing a push lawnmower and tools for which he was sent to prison. This sentence was based largely on his COMPAS score:

> "Zilly and his lawyer agreed to a plea deal with prosecutors, in which the state would recommend one year in a county jail followed by supervision to ensure Zilly would 'stay[] on the right path.' However, Judge James Babler overturned the plea deal and sentenced Zilly to two years in prison, stating: 'When I look at the risk assessment . . . it is about as bad as it could be.' The judge referenced the score generated by COMPAS, which calculated Zilly as high risk for future violent crime and medium risk for general recidivism. In an appeals hearing, *Judge Babler explained his sentencing decision: 'Had I not had the COMPAS, I believe it would likely be that I would have given one year, six months.*'" [4], 319-320,emphasis added).

Clearly, the COMPAS scores are determinative of sentencing decisions in this case. This is just one example of a more general issue, namely that of automation bias and the "technology effect":

> "While it may be common practice to express deference to a judge's discretion, the influence of the "technology effect" deteriorates the trustworthiness of judiciary discretion. With or without a warning label, judges consistently give technology and forensic-based evidence heavier weight than other factors, whether the judges giving such weight realize that they are doing so or not. Studies have shown that

> people have 'automation bias' and, therefore, place their trust in computer-generated assessments even when faced with evidence of the systems' inaccuracies" (Freeman 2017, 97-98).

As Danielle Citron puts it in her discussion of automation bias and technological due process, "Automation bias effectively turns a computer program's suggested answer into a trusted final decision" [5], 1272). Furthermore, given individual's- including judges'- inability to disregard certain factors once they have been exposed to the individual's conscious, the Wisconsin Supreme Court's proposed solution of cautionary statements will do little to ensure that COMPAS risk scores are limited to a merely corroborative role in sentencing decisions.

This raises the question of what evidential role COMPAS scores should play in judicial discretion and the appropriate norms governing how much weight, if any, such evidence should be given when considered alongside other lines of evidence. The different confirmatory values of evidence produced by COMPAS for race-specific predictive hypotheses about rearrest can provide guidance as to how much weight to assign the risk scores within a given body of evidence for a specific hypothesis and judgment. If the confirmatory value of such evidence is significant or greater than some designated threshold, then the added value of the COMPAS risk score as corroborative evidence to other lines of evidence is greater, and the overall body of evidence, including the COMPAS risk score, is more secure. This body of evidence, considered holistically, thus provides stronger, in terms of more reliable, support for the final judgment or decision, than the same body of evidence absent the COMPAS risk score. In other words, the COMPAS evidence can confirm a judgment to a greater degree than otherwise possible given the remaining available lines of evidence.

However, where the COMPAS evidence is weaker, judges should require stronger alternative lines of evidence for a given decision. If such evidence is unavailable, the addition of weak COMPAS evidence cannot strengthen a weak body of existing legal evidence.

Consider the following example of a black defendant. There are three lines of evidence in a case:

> (e_1): the defendant is unemployed and lives in a poor neighborhood;
> (e_2): testimony from vengeful associates that the defendant has an uncontrollable habit of engaging in criminal behavior;
> (e_3): the defendant has a "high risk" COMPAS score.

A prosecutor may grant that each line of evidence (e.g., the testimony of an angry former associate) is weak, but that, taken together with the COMPAS risk score, it is probable that the defendant will reoffend and thus should be incarcerated pending trial.

However, it is an error to use the COMPAS score as an independent line of evidence that is on par with the other lines of evidence, (e_1) and (e_2). This is because, in our example, the COMPAS score depends on the other lines of evidence, like (e_1) concerning the defendant's employment status, independently considered. When the COMPAS score is weighted and used as independent evidence in conjunction with (e_1) and (e_2), employment status is entered as evidence effectively twice and thus given outsized importance. This mistake arises because the COMPAS score is not a third or independent line of evidence for the hypothesis about the likelihood of a defendant reoffending. Rather, a COMPAS score represents a report on the conditional probability that a defendant will reoffend given the existing evidence being independently considered; in this case, (e_1) and (e_2) as well as the system's training data. In this way, a COMPAS score is analogous to an expert opinion about the degree to which a given body of evidence supports a hypothesis. One cannot then conditionalize the probability of a hypothesis on the expert report because the report does not represent new evidence, only a probabilistic report of the available evidence. Thus, it is inappropriate to treat the COMPAS score as an independent piece of evidence on par with the remaining evidence and use such a score to strengthen an existing body of weak evidence. If, given all the available evidence, the posterior probability of a hypothesis is low—e.g., $p(H\backslash((e_1)\&(e_2)\&(e_3)) = 25\%$—and $p(H|\ e_3)$ (representing the COMPAS score of high risk for a black defendant) is low (11.3%), one cannot use the COMPAS score to argue for a different or higher posterior probability such that $p(H|e_3)$ & $p(H\backslash((e_1)\&(e_2)\&(e_3))$ is $> 50\%$.

Moreover, weakly confirmatory COMPAS evidence cannot corroborate or compensate for weak evidence elsewhere, without becoming determinative of the resulting legal judgment. If a judgment is reached through reliance on a variety of weak evidence, including COMPAS evidence, then the decision-maker may be blame-worthy for failing to uphold their epistemic and moral responsibility to make decisions, especially potentially harmful decisions, based on strong and reliable evidence.

Conversely, it is worth noting that no matter how strong the COMPAS score evidence is, it is still probabilistic evidence. It cannot provide certain evidence with the posterior probability of a defendant reoffending $p(H|e_{total}) = 1$. This is important because the probabilistic nature of such evidence can place important limits on the types of decisions and actions that may be warranted based on such evidence alone or when combined with other lines of probabilistic evidence.

Through a confirmational analysis of COMPAS risk scores as evidence, individuals can better delineate the appropriate and actual weight given to such risk scores in

judicial judgment. This can provide guidance as to how decision makers ought to value COMPAS scores as legal evidence within the context of other lines of evidence. Furthermore, an analysis of the confirmational value of COMPAS evidence can indicate whether and how to hold legal decision makers accountable for the misuse of COMPAS scores as determinative or corroborating evidence when it is weakly confirmatory or merely a report on the independently considered evidence, or the failure to appropriately weight COMPAS scores when it is strongly confirmatory for a given predictive hypothesis. For example, given the preceding analysis of the strong confirmatory value of a COMPAS risk score of "low risk" for black defendants, a judge should be required to provide adequate justification for overriding such a score and nonetheless treating a black defendant as "high risk" in pre-sentencing and sentencing decisions such as pre-trial arrest and detention, setting bail at extremely high amounts, or denying a defendant bail. In this way, the confirmational value of COMPAS risk scores can introduce an additional demand for judicial justification of a high-stakes decision.

7 Concluding remarks

Since 2016, the debate surrounding algorithmic fairness in the context of criminal sentencing risk assessment has been ongoing. The widely-used COMPAS system, developed by Northpointe Co., was found to exhibit racial bias, with critics like ProPublica pointing out disparities in error rates, particularly with false positive errors favoring white defendants. This bias in classification can profoundly impact defendants, influencing decisions regarding sentence severity, pre-trial detention, and bail eligibility.

Discussions have revolved around the balance between calibration and error-rate equity in algorithmic predictions. Despite efforts to mitigate bias through technical means, unfairness persists, underscoring the need for conceptual clarity and a new approach to algorithmic fairness.

This paper proposes an alternative framework centered on evaluating the confirmational value of algorithmic outputs for criminal justice decisions. It advocates for ethical guidelines, emphasizing the importance of confirmatory evidence, particularly for high-stakes decisions, and equal thresholds of confirmational significance for sentencing decisions across racial groups.

Given the weakness of COMPAS scores for confirming black defendants' high level recidivism risk, such scores should not be used to justify imposing harsh or harsher sentences on black defendants. Conversely, given the strength of COMPAS evidence for black defendant's low recidivism risk, such evidence should be used to as a mitigating factor in sentencing to reduce sentences for black defendants, support early release, etc. Taken together, decisions based on COMPAS evidence may provide a corrective to the historical and current racialized patterns of arrest, incarceration, and harsh sentencing. Such evidence can be restricted from being used to arrest and imprison more black defendants for longer periods of time and can be used to reduce the pre-trial arrest and incarceration of black defendants. Mitigating the negative outcomes of criminal sentencing for black defendants may reduce existing racial disparities in the U.S. criminal justice system. Moreover, this can be done without violating equal treatment of black and white defendants because both black and white defendants are held to the same evidential threshold of confirmational significance.

By prioritizing equally confirmatory scores over equal scores, this framework seeks to reconcile accuracy and fairness, potentially reducing racial disparities in sentencing outcomes. It suggests leveraging COMPAS evidence appropriately to mitigate biases and promote equitable treatment in the criminal justice system. Thus, this account avoids the problem of the "impossibility of fairness" and the tradeoff between equality of treatment and equality of outcomes.

This approach not only offers theoretical underpinnings but also practical guidance for decision-makers, emphasizing accountability and transparency in the use of algorithmic predictions in sentencing. By adopting a confirmation theoretic perspective, this framework offers a promising avenue for addressing the complexities of algorithmic fairness in criminal justice and advancing the discourse on the "impossibility of fairness" problem in AI ethics.

Author contributions Single-author manuscript.

Data availability Not applicable.

Declarations

Conflict of interest The author has no relevant financial or non-financial interests to disclose.

Informed consent I grant my consent for publication.

References

1. Angwin, J., Jeff, L., Surya, M., Lauren, K.: 2016. Machine bias. In Ethics of data and analytics, pp. 254–264. Auerbach Publications. Accessed March 14, 2023. https://www.propublica.org/article/machine- bias-risk-assessments-in-criminal-sentencing.
2. Beriain Miguel, I.D.: Does the use of risk assessments in sentences respect the right to due process? A critical analysis of the Wisconsin v. Loomis ruling. Law Probab. Risk **17**(1), 45–53 (2018)
3. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: the state of the art. Sociol Methods Res **50**(1), 3–44 (2021). https://doi.org/10.1177/0049124118782533

4. Carlson, A. M.: The need for transparency in the age of predictive sentencing algorithms. Iowa L. Rev.**103**, 303 (2017)
5. Citron, D. K.: Technological due process. Wash. UL Rev. **85**, 1249 (2007)
6. Chouldechova, A.: Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data **5**(2), 153–163 (2017). https://doi.org/10.1089/big.2016.0047
7. Corbett-D., Sam, E.P., Avi F., and Sharad G.: A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. (2016)
8. Costanza-C, Sasha.: Design justice: community-led practices to build the worlds we need. https://doi.org/10.7551/mitpress/12255.001.0001. (2020)
9. Custers, B.: AI in criminal law: an overview of AI applications in substantive and procedural criminal law. Law Artif. Intell. (2022). https://doi.org/10.1007/978-94-6265-523-2_11
10. Dieterich, W, Christina M, and Tim B. COMPAS risk scales: demonstrating accuracy equity and predictive parity. Northpointe Inc. **7**(74), 1 (2016)
11. Flores, A.W., Kristin, B., Lowenkamp, T.C.: False positives, false negatives, and false analyses: a rejoinder to machine bias: there's software used across the country to predict future criminals. and it's biased against blacks. Fed. Probation **80**, 38 (2016)
12. Freeman, K.: Algorithmic injustice: how the Wisconsin supreme court failed to protect due process rights in State v. Loomis. North Carolina J. Law and Technol. **18**(5), 75 (2016)
13. Green, B.: Escaping the impossibility of fairness: from formal to substantive algorithmic fairness. Philos. Technol. **35**(4), 90 (2022). https://doi.org/10.1007/s13347-022-00584-6
14. Humerick, JD.: Reprogramming fairness: affirmative action in algorithmic criminal sentencing. (2020)
15. Huq, A.: Racial equity in algorithmic criminal justice. Duke Law J. **68**. (2019)
16. Isaac, W.S.: Hope, hype, and fear: the promise and potential pitfalls of artificial intelligence in criminal justice. Ohio St. J. Crim. L. **15**, 543 (2017)
17. Joyce, J.: Bayes' theorem. The Stanford encyclopedia of philosophy (Fall 2021 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2021/entries/bayes-theorem/
18. Kleinberg, J., Mullainathan, S. and Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. (2016). arXiv. http://arxiv.org/abs/1609.05807.
19. Larson, J., Surya M., Lauren Kirchner and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. Accessed March 14, 2023. (2016). https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
20. Lightbourne, J.: Damned lies and criminal sentencing using evidence-based tools. Duke L. Tech. Rev. **15**, 327 (2016)
21. Liptak, A.: Sent to Prison by a Software Program's Secret Algorithms. The New York Times, May 1, (2017), sec. U.S. https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html.
22. Abdul, M.M.: Criminal courts' artificial intelligence: the way it reinforces bias and discrimination. AI and Ethics **2**(1), 233–245 (2022)
23. Mayson, S.G.: Bias in, bias out. YAle lJ. **128**, 2218 (2018)
24. McKay, C.: Predicting risk in criminal procedure: actuarial tools, algorithms, AI and judicial decision-making. Curr. Issues Crim. Just. **32**(1), 22–39 (2020). https://doi.org/10.1080/10345329.2019.1658694
25. Park, AL.: Injustice ex machina: Predictive algorithms in criminal sentencing.UCLA Law Rev.**19** (2019).
26. Rudner, R.: The scientist Qua scientist makes value judgments. Philos. Sci. **20**(1), 1–6 (1953)
27. Starr, S.B.: Evidence-based sentencing and the scientific rationalization of discrimination. Stan. L. Rev. **66**, 803 (2014)
28. Thomson, J.J.: Liability and individualized evidence. Law Contemp. Probs. **49**, 199–219 (1986)
29. Wisser, L.: Pandora's algorithmic black box: the challenges of using algorithmic risk assessments in sentencing. AM. Crim. L. Rev. **56**, 1811 (2019)