11

Supervising Automated Decisions

Tatiana Cutts

11.1 INTRODUCTION

AI and ADM tools can help us to make predictions in situations of uncertainty, such as how a patient will respond to treatment, and what will happen if they do not receive it; how an employee or would-be employee will perform; or whether a defendant is likely to commit another crime. These predictions are used to inform a range of significant decisions about who should bear some burden for the sake of some broader social good, such as the relative priority of organ transplant amongst patients; whether to hire a candidate or fire an existing employee; or how a defendant should be sentenced.

Humans play a critical role in setting parameters, designing, and testing these tools. And if the final decision is not purely predictive, a human decision-maker must use the algorithmic output to reach a conclusion. But courts have concluded that humans also play a corrective role[1] – that, even if there are concerns about the predictive assessment, applying human discretion to the predictive task is both a necessary and sufficient safeguard against unjust ADM.[2] Thus, the focus in academic, judicial, and legislative spheres has been on making sure that humans are equipped and willing to wield this ultimate decision-making power.[3]

I argue that this focus is misplaced. Human supervision can help to ensure that AI and ADM tools are fit for purpose, but it cannot make up for the use of AI and ADM tools that are not. Safeguarding requires gatekeeping – using these tools just when we can show that they take the right considerations into account in the right way. In this chapter, I make some concrete recommendations about how to determine

[1] See e.g. *State v Loomis* 881 N.W.2d 749 (Wis. 2016) at [71].
[2] Ibid at [92].
[3] See e.g. Reuben Binns, 'Algorithmic Decision-Making: A Guide for Lawyers' (2020) 25 *Judicial Review* 2, 6.

https://doi.org/10.1017/9781009334297.016 Published online by Cambridge University Press

whether AI and ADM tools meet this threshold, and what we should do once we know.

11.2 THE DETERMINATIVE FACTOR

In 2013, Eric Loomis was convicted of two charges relating to a drive-by shooting in La Crosse, Wisconsin: 'attempting to flee a traffic officer and operating a motor vehicle without the owner's consent'.[4] The pre-sentence investigation (PSI) included COMPAS risk and needs assessments.[5] COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a suite of ADM/AI tools developed and owned by Equivant.[6] These tools are designed to predict recidivism risk for individual offenders and patterns across wider populations, by relying upon inferences drawn from representative pools of data. The sentencing judge explicitly invoked each COMPAS assessment to justify a sentence of six years in prison and five years of extended supervision.[7]

Though the literature often refers to 'the COMPAS algorithm',[8] COMPAS is not a single algorithm that produces a single risk-score; rather, the COMPAS software includes a range of ADM tools that use algorithms to predict risk, which are described by Equivant as 'configurable for the user'.[9] The tools available include: Pre-Trial Services,[10] which principally concern the risk that the accused will flee the jurisdiction; and three assessments (the General Recidivism Risk scale (GRR), the Violent Recidivism Risk scale (VRR), and the 'full assessment') which involve predictions about recidivism. The GRR, VRR, and full assessment are designed to inform public safety considerations that feed into decisions about resource-allocation across populations,[11] and are used in several jurisdictions to decide how to treat individual offenders.

[4] See e.g. Brief of Defendant-Appellant, *State v Loomis*, No 2015AP157-CR (Wis Ct App 2015), 2015 WL 1724741, 1–3; *State v Loomis* at 754.

[5] See e.g. '*State v Loomis*: Wisconsin Supreme Court Requires Warning before Use of Algorithmic Risk Assessments in Sentencing' (2017) 130 *Harvard Law Review* 1530.

[6] Previously Northpointe.

[7] See e.g. Brief of Defendant-Appellant, *State v Loomis*, 9.

[8] See e.g. Ellora Israni, 'Algorithmic due Process: Mistaken Accountability and Attribution in *State v Loomis*' (31 August 2017) *JOLT Digest* <https://jolt.law.harvard.edu/digest/algorithmic-due-process-mistaken-accountability-and-attribution-in-state-v-loomis-1> (accessed 25 August 2022); Leah Wisser, 'Pandora's Algorithmic Black Box: The Challenges of Using Algorithmic Risk Assessments in Sentencing' (2019) 56 *American Criminal Law Review* 1811.

[9] Northpointe Institute for Public Management, *Measurement and Treatment Implications of COMPAS Core Scales* (Report, 30 March 2009) 4.

[10] Ibid.

[11] See generally ibid and Equivant, *Practitioner's Guide to COMPAS Core* (2019). Dr David Thompson testified at the post-conviction hearing, telling the court that COMPAS was originally designed to help corrections allocate resources and to identify individual needs in the community. Brief of Defendant-Appellant, *State v Loomis*, 13.

As the COMPAS software is a trade secret, only the score is revealed to the defendant and court. Nevertheless, Equivant's public materials explain that the GRR includes factors such as: 'criminal associates';[12] 'early indicators of juvenile delinquency problems';[13] 'vocational/educational problems';[14] history of drug use;[15] and age.[16] The enquiry into 'vocational/educational problems' in turn includes data points that are identified by defendants' responses to questions such as: 'how hard is it for you to find a job above minimum wage'; 'what were your usual grades in school'; and 'do you currently have a skill, trade, or profession at which you usually find work'.[17] Equivant notes that these data points are strongly correlated to 'unstable residence and poverty', as part of a pattern of 'social marginalisation'.[18]

The 'full assessment'[19] is designed to assess a much wider set of 'criminogenic need'[20] factors, which are identified by the literature as 'predictors of adult offender recidivism'.[21] These include 'anti-social friends and associates';[22] poor family and/or marital relationships (including whether the defendant was raised by their biological parents, parental divorce or separation, and family involvement in criminal activity, drugs, or alcohol abuse);[23] employment status and prospects;[24] school performance;[25] and 'poor use of leisure and/or recreational time'.[26]

Some of these factors are assessed according to the defendant's own input to a pre-trial questionnaire, some are subjective observations made by the assessing agent, and some are objective data (such as criminal record). Scores are then incorporated by the agent into an overall narrative, which forms the basis of a sentencing recommendation by the district attorney. COMPAS is used by corrections departments, lawyers, and courts across the United States to inform many elements of the criminal process, including decisions about pre-trial plea negotiations; 'jail

[12] Northpointe Institute for Public Management, *Measurement and Treatment Implications of COMPAS Core Scales*, 6.
[13] Ibid.
[14] Ibid.
[15] Ibid.
[16] Ibid, 31.
[17] Northpointe Institute for Public Management, *Measurement and Treatment Implications of COMPAS Core Scales*, 21.
[18] Equivant, *Practitioner's Guide to COMPAS Core*, 31, 45, 57.
[19] Northpointe Institute for Public Management, *Measurement and Treatment Implications of COMPAS Core Scales*, 4.
[20] Equivant, *Practitioner's Guide to COMPAS Core*, 36ff.
[21] Paul Gendreau, Tracy Little, and Claire Goggin, 'A Meta-Analysis of the Predictors of Adult Offender Recidivism: What Works!' (1996) 34 *Criminology* 575.
[22] Equivant, *Practitioner's Guide to COMPAS Core*, 36.
[23] Northpointe Institute for Public Management, *Measurement and Treatment Implications of COMPAS Core Scales*, 13.
[24] Ibid, 22.
[25] Ibid, 23.
[26] Northpointe Institute for Public Management, *Measurement and Treatment Implications of COMPAS Core Scales*, 44.

programming' requirements; community referrals; bail applications; sentencing, supervision, and probation recommendations; and the frequency and nature of post-release contact.[27]

Loomis' PSI included both risk scores and a full criminogenic assessment, and each assessment informed the trial court's conclusion that the 'high risk and the high needs of the defendant' warranted a six-year prison sentence with extended supervision.[28] Loomis filed a motion for post-conviction relief, arguing that the court's reliance on COMPAS violated his 'due process' rights in three ways: first, Loomis argues that 'the proprietary nature of COMPAS' prevented him from assessing the accuracy of predictive determinations;[29] second, Loomis argued that use of COMPAS denied him the right to an 'individualized' sentence;[30] finally, he argued that COMPAS 'improperly uses gendered assessments'.[31] The trial court denied the post-conviction motion, and the Wisconsin Court of Appeals certified the appeal to the Supreme Court of Wisconsin (SCW).

Giving the majority judgment, Ann Walsh Bradley J. rejected the claim that Loomis had a right to see the internal workings of the COMPAS algorithms; it was, she said, enough that the statistical accuracy of the COMPAS risk scales had been verified by external studies,[32] and that Loomis had access to his own survey responses and COMPAS output.[33] She noted that 'some studies of COMPAS risk assessment have raised questions about whether they disproportionality classify minority offenders as having a higher risk of recidivism'.[34] Nevertheless, the judge felt that this risk could be mitigated by requiring that the sentencing court be provided with an explanatory statement outlining possible shortcomings in overall risk prediction and the distribution of error.[35]

Addressing Loomis' argument that use of the COMPAS scores infringed his right to an 'individualized' sentence, the judge considered that '[i]f a COMPAS risk assessment were the determinative factor considered at sentencing this would raise due process challenges regarding whether a defendant received an individualized sentence'.[36] By contrast, 'a COMPAS risk assessment may be used to enhance a judge's evaluation, weighing, and application of the other sentencing evidence in

[27] See generally State of Wisconsin Department of Corrections, *State of Wisconsin Department of Corrections Electronic Case Reference Manual* (Web Page) <https://doc.helpdocsonline.com/arrest-and-adjudication> (accessed 25 August 2022). For instance, Wisconsin DOC recommends that probation be imposed if one of the 'eight criminogenic needs' identified by COMPAS is present.

[28] See e.g. Brief of Defendant-Appellant, *State v Loomis*, 10.

[29] *State v Loomis* at [6].

[30] Ibid at [34].

[31] Ibid.

[32] Ibid at [58].

[33] Ibid at [55].

[34] Ibid at [61], [100].

[35] Ibid at [100].

[36] Ibid at [104], [120].

the formulation of an individualized sentencing program appropriate for each defendant',[37] as 'one tool available to a court at the time of sentencing'.[38] The judge emphasised that the court, like probation officers, should feel empowered to disagree with algorithmic predictions as and where necessary.[39]

Finally, the judge rejected Loomis' arguments about the 'inappropriate' use of gendered assessments, noting that 'both parties appear to agree that there is statistical evidence that men, on average, have higher recidivism and violent crime rates compared to women'.[40] Indeed, the judge concluded that 'any risk assessment which fails to differentiate between men and women will misclassify both genders'.[41]

Applying these considerations to the instant case, the judge concluded that there had been no failure of due process, because the COMPAS score had been 'used properly'.[42] Specifically, 'the circuit court explained that its consideration of the COMPAS risk scores was supported by other independent factors, its use was not determinative in deciding whether Loomis could be supervised safely and effectively in the community'.[43]

Human reasoning clearly feeds into processes of AI design and development, and humans are often needed to use the predictive outputs of algorithmic processes to make decisions. The question is whether the SCW was correct to conclude that, even if there are doubts about the quality of the algorithmic assessment (overall accuracy, distribution of the risk of error, or some other concern), human supervision at the time of decision-making is a sufficient safeguard against unjust decisions.

11.3 INDIVIDUALISM AND RELEVANCE

Justice is sometimes described as an 'individualistic' exercise, concerned with the 'assessment of individual outcomes by individualized criteria'.[44] Prima facie, this seems to be a poor fit use of statistics to make decisions about how to treat others. As a science, 'statistics' is the practice of amassing numerical data about a subset of some wider population or group, for the purpose of inferring conclusions from the former about the latter. And in Scanlon's words, 'statistical facts about the group to which a person belongs do not always have the relevant justificatory force'.[45]

[37] Ibid at [92]; *Malenchik v State* (2010) Ind 928 NE 2d 564, 573, emphasis added.
[38] Ibid; *State v Samsa* 359 (2015) Wis 2d 580 at [13], emphasis added.
[39] Ibid at [71]. Wisconsin Department of Corrections guidance states that 'staff should be encouraged to use their professional judgment and override the computed risk as appropriate' (n 27).
[40] Ibid at [78].
[41] Ibid at [83].
[42] Ibid at [104].
[43] Ibid at [9].
[44] J Waldron, 'The Primacy of Justice' (2003) 9 *Legal Theory* 269, 284.
[45] See e.g. TM Scanlon, *Why Does Inequality Matter?* (Oxford University Press, 2017) 27.

But we often make just decisions by reference to the characteristics of a group to which the decision-subject belongs. During the COVID-19 pandemic, decisions about how to prioritise vaccination and treatment were made by governments and doctors across the world on the basis of facts about individuals that were shared with a representative sample of the wider population. There being statistical evidence to demonstrate that those with respiratory or auto-immune conditions were at an aggravated risk of serious harm, patients with these conditions were often prioritised for vaccination, whilst mechanical ventilation was reserved for seriously ill patients who were likely to survive treatment.[46] Making 'individualised' decisions does not require us to ignore relevant information about other people; it simply requires us *not* to ignore relevant information about the decision-subject.

In this context, 'relevant' means rationally related to the social goal of improving health outcomes. A doctor ought to consider features of particular patients' circumstances that shape their needs and likely treatment outcomes. She might, for instance, decide to ventilate an older but healthy patient – taking into account the patient's age and an assessment of their overall well-being to conclude that treatment survival is highly likely. This is an 'individualised' assessment, in that it takes into account relevant facts, which are characteristics that this patient shares with others. By contrast, her decision should be unaffected by facts that do not bear on treatment success, such as whether the patient is a family member.

So, to justify a policy that imposes a burden on some people for the sake of a social goal, the policy must aim at some justified social goal, to which our selection criteria must be rationally related. The next question is whether ADM and AI tools can help us to make decisions on the basis of (all and only) relevant criteria.

11.4 STATISTICAL RULES AND RELEVANCE

In 1943, Sarbin published the results of a study comparing the success of 'actuarial' (statistical) and 'clinical' (discretionary) methods of making predictions.[47] The goal of the exercise was to determine which method would predict academic achievement more accurately. To conduct the experiment, Sarbin chose a sample of 162 college freshman, and recorded honor-point ratios at the end of the first quarter of their freshman year.[48]

Actuarial assessments were limited and basic: they were made by entering two variables (high school percentile rank and score on college aptitude test) into a two-variable regression equation. Individual assessments were made by the university's

[46] See e.g. British Medical Association, *COVID-19 Ethical Issues: A Guidance Note* (Report, 2020) <www.bma.org.uk/media/2226/bma-covid-19-ethics-guidance.pdf> (accessed 25 August 2022).

[47] Theodore R Sarbin, 'A Contribution to the Study of Actuarial and Individual Methods of Prediction' (1943) 48 *American Journal of Sociology* 593.

[48] The ratio of credits to grades that have been converted into honour points.

clinical counsellors and included a far broader range of variables: an interviewer's form and impressions; test scores for aptitude, achievement, vocation, and personality; and the counsellor's own impressions.

Sarbin found that the actuarial method was more successful by a small margin than the individual method at predicting academic achievement, concluding that 'any jury sitting in judgment on the case of the clinical versus the actuarial methods must on the basis of efficiency and economy declare overwhelmingly in favour of the statistical method for predicting academic achievement'.[49]

Many other studies have produced similar results across a range of different areas of decision-making, including healthcare, employee performance, and recidivism.[50] Conrad and Satter compared statistical and discretionary predictions about the success of naval trainees in an electrician's mate school.[51] They pitted the output of a two-factor regression equation (electrical knowledge and arithmetic reasoning test scores) against the predictions of interviewers on the basis of test scores, personal history data, and interview impressions. Their conclusions favoured the statistical method.

In principle, human reasoning that is unconstrained by (statistical or other) rules can be sensitive to a limitless range of relevant facts. But there are several caveats to this promising start. First, humans are easily influenced by irrelevant factors, or over-influenced by relevant factors, and extremely poor at recognising when we have been influenced in this way. There is now a great deal of literature detailing the many 'cognitive biases' that affect our decision-making, such as: 'illusory correlation' (hallucinating patterns from a paucity of available data) and 'causal thinking' (attributing causal explanations to those events).[52]

Second, the availability of more information does not necessarily translate into a broad decision process. Indeed, Sarbin found that the high-school rank and college aptitude test accounted for 31 per cent of the variance in honour-point ratio and for 49 per cent in the clinical predictions in his experiment[53] – which is to say, the counsellors *overweighted* these two factors, and did not take into account any other measures available to them in a systematic way.

Thus, this theoretical advantage often fails to translate into better decision-making. Yet, AI and ADM tools are no panacea for decision-making under conditions of uncertainty. Predictive success depends on many factors, one of which is the relationship between the chosen proxy and the social goal in question. Sarbin himself noted the limitations of using honour-point ratio as a proxy for academic

[49] Sarbin, 'A Contribution to the Study of Actuarial and Individual Methods of Prediction', 600.
[50] See e.g. Daniel Kahneman, *Thinking Fast and Slow* (Penguin, 2011) 222.
[51] HS Conrad and GA Satter, *Use of Test Scores and Quality Classification Ratings in Predicting Success in Electrician's Mates School* (Office of Social Research and Development Report No 5667, September 1945).
[52] See e.g. Kahneman, *Thinking Fast and Slow*, 77, 115.
[53] Sarbin, 'A Contribution to the Study of Actuarial and Individual Methods of Prediction', 596.

achievement,[54] and the same concerns arise in many other areas of decision-making. For instance, predictions about recidivism are hampered by the fact that crime reports, arrest, and conviction data poorly mirror the actual incidence of crime.

Predictive success also depends upon the quality of the data, including whether that data is representative of the wider target population. The anti-coagulant medication warfarin is regularly prescribed to patients on the basis of dosing algorithms, which incorporate race as a predictor along with clinical and genetic factors.[55] Yet, most of the studies used to develop these algorithms were conducted in cohorts with >95 per cent white European ancestry, and there is now robust evidence that these algorithms assign a 'lower-than-needed dose' to black patients, putting them at serious risk of heart attack, stroke, and pulmonary embolism.[56]

The Model for End-Stage Liver Disease (MELD) is used to calculate pre-treatment survival rates in liver transplant patients, on the basis of factors such as levels of bilirubin and creatinine in the blood. MELD scores are used to make decisions about which patients to prioritise for transplant. Yet, the MELD was developed on the basis of several studies that either did not report sex data, or which reported a statistical makeup of 70 per cent men (without disaggregating data in either case),[57] and a recent study has found that women have a 19 per cent increased risk of wait-list mortality compared to men with the same MELD scores.[58]

So, AI and ADM tools can sometimes help us to make decisions on the basis of criteria that are rationally related to our social goal. Whether they do have this effect depends (inter alia) upon the quality of the data and the relationship between the chosen proxy and social goal in question. Yet, there may be countervailing reasons to *exclude* certain relevant factors from the decision-making process. I turn to these considerations now.

11.5 CHOICE

Overdose deaths from opioids across the United States increased to 75,673 in the twelve-month period ending in April 2021, up from 56,064 the year

[54] Ibid, 594.

[55] 'Race-Specific Dosing Guidelines Urged for Warfarin' (February 2017) *Ash Clinical News* <https://ashpublications.org/ashclinicalnews/news/2145/Race-Specific-Dosing-Guidelines-Urged-for-Warfarin> (accessed 25 August 2022).

[56] Nita A Limdi et al, Race Influences Warfarin Dose Changes Associated with Genetic Factors (2015) 126 *Blood* 539, 544.

[57] See e.g. Russell Wiesner et al, 'Model for End-Stage Liver Disease (MELD) and Allocation of Donor Livers' (2003) 124 *Clinical-Liver, Pancreas, and Biliary Tract*; B Brandsaeter et al, 'Outcome Following Liver Transplantation for Primary Sclerosing Cholangitis in the Nordic Countries' (2003) 38 *Scandinavian Journal of Gastroenterology* 1176.

[58] CA Moylan et al, 'Disparities in Liver Transplantation before and after Introduction of the MELD Score' (2008) 300 *JAMA* 2371.

before.[59] In 2020, more people in San Francisco died of opioid overdoses than of COVID-19.[60] A significant portion of that uptick has been attributed to a pattern of aggressive and successful marketing of the prescription opioid OxyContin between 1996 and 2010. When OxyContin was reformulated in 2010 to make it more difficult to abuse, many of those who were addicted to prescription opioids switched to heroin and, eventually, fentanyl. One study found that 77 per cent of individuals who used both heroin and nonmedical pain relievers between 2011 and 2014 had initiated their drug use with prescription opioids,[61] and there is now a broad consensus that the introduction of OxyContin can 'explain a substantial share of overdose deaths' over twenty years.[62]

Many different measures have been taken to prevent addiction and abuse, and to support those who are suffering from addiction. One preventative measure is the Opioid Risk Tool (ORT), which was published in 2005 on the basis of several studies that identified correlations between certain facts and opioid misuse.[63] This questionnaire, which is used in several jurisdictions across the world, consists of ten scorable components, including family or personal history of substance abuse or psychological disorder; patient age; and (if the patient is female) a history of preadolescent sexual abuse.

According to Webster, author of the ORT, his goal was 'to help doctors identify patients who might require more careful observation during treatment, not to deny the person access to opioids'.[64] Yet, the ORT is in fact used in clinical practice to decide whether to deny or withdraw medical treatment from patients,[65] which has had a severe impact on patients, particularly women, who suffer from severe and chronic pain.[66] High ORT scores have resulted in the termination of doctor–patient

[59] See e.g. Centres for Disease Control and Prevention, National Center for Health Statistics, *Drug Overdose Deaths in the US Top 100,000 Annually* (Report, 17 November 2021) <www.cdc.gov/nchs/pressroom/nchs_press_releases/2021/20211117.htm> (accessed 25 August 2022).

[60] See e.g. 'Last Year, More People in San Francisco Died of Overdoses Than of Covid-19' (15 May 2021) *The Economist* <www.economist.com/united-states/2021/05/15/last-year-more-people-in-san-francisco-died-of-overdoses-than-of-covid-19> (accessed 25 August 2022).

[61] Pradip K Muhuri, Joseph C Gfroerer, and M Christine Davies, 'Associations of Nonmedical Pain Reliever Use and Initiation of Heroin Use in the United States' (2013) *CBHSQ Data Review* <www.samhsa.gov/data/sites/default/files/DR006/DR006/nonmedical-pain-reliever-use-2013.htm> (accessed 25 August 2022).

[62] 'Patrick Radden Keefe Traces the Roots of America's Opioid Epidemic' (13 May 2021) *The Economist* <www.economist.com/books-and-arts/2021/05/13/patrick-radden-keefe-traces-the-roots-of-americas-opioid-epidemic> (accessed 25 August 2022).

[63] Lynn R Webster and Rebecca M Webster, 'Predicting Aberrant Behaviors in Opioid-Treated Patients: Preliminary Validation of the Opioid Risk Tool' (2005) 6 *Pain Medicine* 432.

[64] Lynn Webster, 'Another Look at the Opioid Risk Tool' (29 June 2022) *Pain News Network* <www.painnewsnetwork.org/stories/2022/6/29/another-look-at-the-opioid-risk-tool> (accessed 25 August 2022).

[65] See e.g. NR Brott, E Peterson, and M Cascella, *Opioid Risk Tool* (StatPearls Publishing, 2022).

[66] Jennifer D Oliva, 'Dosing Discrimination: Regulating PDMP Risk Scores' (2022) 110 *California Law Review* 47.

https://doi.org/10.1017/9781009334297.016 Published online by Cambridge University Press

relationships, as well as attracting negative interpersonal treatment by members of medical staff, adding emotional distress to physical pain.[67]

Many authors have objected to use of the ORT to make prescribing decisions on the basis that this practice discriminates against women.[68] Yet, 'discrimination' is an umbrella term. The wrongfulness of discrimination lies in the fact that the characteristics upon which we make decisions that disadvantage certain groups do not justify that treatment,[69] and there are different reasons to object to policies that have this effect.

The first reason that we might invoke to object to decision-making policies or practices that rely upon the ORT is that our decisions are based on criteria (such as the preadolescent sexual abuse of women) that are not rationally related to the social goal of preventing and reducing opioid addiction.[70] The second reason concerns the broader significance of this failure to develop and implement sound medical policy. It might, for instance, indicate that policymakers have taken insufficient care to investigate the connection between the sexual abuse of women and opioid abuse. When the consequence is placing the risk of predictive error solely upon women, the result is a failure to show equal concern for the interests of all citizens.[71] Finally, we might object to use of the ORT on the basis that the policy reflects a system in which women are treated as having a lower status than men – a system in which practices of exclusion are stable, so that women are generally denied opportunities for no good reason.[72]

But there is also an objection to policies that rely upon the ORT that has nothing to do with inequality. The argument is that, when we impose burdens on some people for the sake of some benefit to others, we should (wherever possible) give those people the opportunity to avoid those burdens by choosing appropriately. Policies that impose burdens upon individuals on the basis of facts about the actions

[67] Maia Szalavitz, 'The Pain Was Unbearable. So W Did Doctors Turn Her Away?' (11 August 2022) *Wired* <www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/> (accessed 25 August 2022).

[68] Oliva, 'Dosing Discrimination: Regulating PDMP Risk Scores'.

[69] See e.g. Scanlon, *Why Does Inequality Matter?*, 26.

[70] See e.g. Constanza Daigre et al, 'History of Sexual, Emotional or Physical Abuse and Psychiatric Comorbidity in Substance-Dependent Patients' (2015) 2293 *Psychiatry Research* 43.

[71] On equal concern generally, see Scanlon, *Why Does Inequality Matter?*, ch. 2.

[72] This is often what we mean when we talk about discrimination, and Webster makes an allegation of this sort when he says: 'the ORT has been weaponized by doctors who are looking for a reason to deny patients – particularly, women – adequate pain medication'; Lynn R Webster and Rebecca M Webster, 'Predicting Aberrant Behaviors in Opioid-Treated Patients: Preliminary Validation of the Opioid Risk Tool'. See also 'The Opioid Risk Tool Has Been Weaponized against Patients' (21 September 2019) *Pain News Network* <www.painnewsnetwork.org/stories/2019/9/21/the-opioid-risk-tool-has-been-weaponized-against-pain-patients> (accessed 25 August 2022).

of others, such as sexual abuse and patterns of family drug abuse, deny those opportunities.

Take the following hypothetical, which I adapt from Scanlon's *What We Owe to Each Other*:[73]

> *Hazardous Waste*: hazardous waste has been identified within a city's most populous residential district. Moving the waste will put residents at risk by releasing some chemicals into the air. However, leaving the waste in place, where it will seep into the water supply, creates a much greater risk of harm. So, city officials decide to take the necessary steps to move and dispose of the waste as safely as possible.

City officials have an important social goal, of keeping people safe. That goal involves the creation of a 'zone of danger' – a sphere of activity that residents cannot perform without serious risk of harm. Accordingly, to justify such a policy, officials need to take precautions that put people in a sufficiently good position to take actions to avoid suffering the harm. They should fence the sites and warn people to stay indoors and away from the excavation site – perhaps by using posters, mainstream media, or text message alerts.

Scanlon uses this hypothetical to explore the justification for the substantive burdens imposed by criminal punishment.[74] There is an important social goal – keeping us safe. The strategy for attaining this goal entails imposing a burden – denying that person some privilege, perhaps even their liberty. Thus, there is now a zone into which people cannot go (certain activities that they cannot perform) without risk of danger. To justify a policy of deliberately inflicting harm on some people, we should give those people a meaningful opportunity to avoid incurring that burden, which includes communicating the rules and consequences of transgression, and providing opportunities for people to live a meaningful life without transgression.

We can apply this logic to the ORT. The ORT was created with an important social goal in mind: preventing opioid misuse and addiction. A zone of danger is created to further that goal: certain patients are denied opioids, which includes withdrawing treatment from those already receiving pain medication, and may include terminating doctor–patient relationships. Patients may also suffer the burden of negative attitudes by medical staff, which may cause emotional suffering and/or negative self-perception. Yet, this time, the patient has no opportunity to avoid the burden of treatment withdrawal: that decision is made on the basis of facts about the actions of others, such as the decision-subject's experience of sexual abuse and/or a family history of drug abuse.

The question, then, is how human oversight bears on these goals: first, making sure that decisions about how to impose burdens on certain individuals for the sake

[73] TM Scanlon, *What We Owe to Each Other* (Harvard University Press, 1998), 256ff.
[74] Ibid, 263ff.

of some social good take into account all and only relevant facts about those individuals; second, making sure that our decisions do not rely upon factors that (even if relevant) we have reason to exclude. In the rest of this chapter, I will look at the knowledge that we need to assess algorithmic predictions, and the threshold against which we make that assessment. I argue that those elements differ markedly according to whether the prediction in question is used to supply information about what a particular decision-subject will do in the future.

11.6 GROUP ONE: PREDICTIONS ABOUT FACTS OTHER THAN WHAT THE DECISION-SUBJECT WILL DO

The first set of cases are those in which the predictive question is about the (current or future) presence of something other than the actions of the decision-subject, such as: the success of a particular course of medical treatment, or the patient's chances of survival without it; social need and the effectiveness of public resourcing; and forensic assessments (e.g., serology or DNA matching). To know whether we are justified in relying upon the predictive outputs of AI and ADM tools in this category, we need to determine whether the algorithmic prediction is more or less accurate that unaided human assessment, and how the risk of error is distributed amongst members of the population.

There are three modes of assessing AI and ADM tools that we might usefully distinguish. The first we can call 'technical', which involves understanding the mechanics of the AI/ADM tool, or 'opening the black box'. The second is a statistical assessment: we apply the algorithm to a predictive task across a range of data, and record overall success and distribution of error. The final mode of assessment is normative: it involves identifying reasons for predictive outputs, by exploring different counterfactuals to determine which facts informed the prediction.

To perform the second and third modes of assessment, we do not need to 'open the black box': the second can be performed by applying the algorithm to data and recording its performance; the third can be performed by applying the algorithm to data and incrementally adjusting the inputs to identify whether and how that change affects the prediction.[75]

To know whether the AI/ADM tool performs better than unaided human discretion, we must perform a statistical assessment. We need not perform either the first or third mode of assessment: we do not need to know the internal workings of the algorithm,[76] and we do not need to know the reasons for the prediction.

[75] See Sandra Wachter, Brent Mittelstadt, and Chris Russell, 'Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR' (2018) 31 *Harvard Journal of Law and Technology* 841.

[76] As Raz puts it, 'Sometimes we can tell that we or others are good at judging matters of a certain kind by the results of our judgements. That would suggest that we, or they, should be trusted

TrueAllele, developed by Cybergenetics and launched in 1994, is an ADM tool that can process complex mixtures of DNA (DNA from multiple sources, in unknown proportions). Prior to the development of sophisticated AI/ADM tools, human discretion was required to process mixtures of DNA (unlike single-source samples), with poor predictive accuracy.[77] Probabilistic genotyping is the next step in forensic DNA, replacing human reasoning with algorithmic processing.

Like COMPAS, the TrueAllele software is proprietary.[78] In *Commonwealth v Foley*,[79] which concerned the defendant's appeal against a murder conviction, one question amongst others was whether this obstacle to accessing the code itself rendered TrueAllele evidence inadmissible in court. On appeal, the defendant argued that the trial court had erred in admitting the testimony of one Dr Mark Perlin, an expert witness for the prosecution, who had communicated the results of a TrueAllele assessment to the Court.

In *Foley*, a sample containing DNA from the victim and another unknown person was found underneath the fingernail of the victim. The mixed sample was tested in a lab, and Perlin testified that the probability that this unknown person was someone other than the defendant was 1 in 189 billion.[80] The defendant argued that the testimony should be excluded because 'no outside scientist can replicate or validate Dr Perlin's methodology because his computer software is proprietary'.[81] On appeal, the Court concluded that this argument 'is misleading because scientists can validate the reliability of a computerized process even if the "source code" underlying that process is not available to the public'.[82]

The TrueAllele prediction is not about what the defendant has done; assessments of guilt or innocence are assessments that the Court (official or jury) must make. Rather, it is about the likelihood of a DNA match – specifically, that the unknown contributor to the DNA sample was someone other than the defendant. In this category of case, I have argued that the Court was correct to indicate that a statistical assessment is sufficient – if such an assessment is sufficiently robust.[83]

If the statistical assessment reveals a rate and distribution of predictive success that is equal to or better than unaided human decision-making, we can justify using the

even when they cannot explain their judgements'. 'This is especially so', he says, 'when understanding of matters in that area is slight'. Joseph Raz, *Engaging Reason: On the Theory of Value and Action* (Oxford University Press, 2002), 246.

77 Katherine Kwong, 'The Algorithm Says You Did It: The Use of Black Box Algorithms to Analyse Complex DNA Evidence' (2017) 31 *Harvard Journal of Law & Technology* 275, 278.

78 Though that may be changing: *'People v H.K'*, *Justia US Law* (Web Page) <https://law.justia.com/cases/new-york/other-courts/2020/2020-ny-slip-op-50709-u.html>.

79 *Commonwealth v Foley* 38 A 3d 882 (PA Super Ct 2012).

80 Ibid, 887.

81 Ibid, 888–89.

82 Ibid.

83 This ought to require assessment by independent entities – entities other than the owner/developer of the algorithm. See e.g. Kwong, 'The Algorithm Says You Did It: The Use of Black Box Algorithms to Analyse Complex DNA Evidence'.

https://doi.org/10.1017/9781009334297.016 Published online by Cambridge University Press

prediction to make decisions. And if it is, we should do consistently, resisting the urge to apply our own discretion to predictions. Of course, we will often take into account the margin of error when applying our judgement to the algorithmic output. For instance, the TrueAllele assessment is only 97 per cent accurate, this ought to affect the weight that we assign to that output in drawing a conclusion about guilt or innocence. But that is a very different exercise from using human judgement to determine the probability of a DNA match in the first place.

If, by contrast, the statistical assessment reveals a rate and distribution of predictive success that is worse than unaided human decision-making, we cannot justify using the prediction to make decisions; there is no meaningful sense in which individual decision-makers can compensate for predictive flaws on an ad hoc basis, and no reason to try, given the availability of a better alternative.

In *Loomis*, the SCW concluded that wrinkles in the COMPAS assessment process and output could be remedied by the application of discretion: '[j]ust as corrections staff should disregard risk scores that are inconsistent with other factors, we expect that circuit courts will exercise discretion when assessing a COMPAS risk score with respect to each individual defendant'.[84] This, I have argued, is an unhappy compromise: either the AI/ADM tool has a better rate and distribution of error, in which case we should not be tempted to override the prediction by applying a clinical assessment, or the AI/ADM tool has a worse rate and distribution of error, in which case unaided human decision-making should prevail unless and until a comprehensive and systematic effort can be made to revise the relevant algorithm.

11.7 GROUP TWO: PREDICTIONS ABOUT WHAT THE DECISION-SUBJECT WILL DO

The second type of case involves the use of AI and ADM tools to make predictive assessments about what the decision-subject will do. This includes, for instance, whether they will misuse drugs or commit a crime, how they will perform on an assessment, or whether they will be a good employee or adoptive parent. To assess whether we are justified in using the predictive outputs of this category of AI and ADM tool, we need to know the facts upon which the prediction is based. This requires us to conduct a counterfactual assessment.

If the prediction is based only on facts that relate to the past actions of the decision-subject, and if the decision-subject has been given a meaningful opportunity to avoid incurring the burden, we may be justified in using the outputs to inform decisions. Whether we are will turn also on the same assessment that we made above: statistical accuracy and the distribution of error. But if the algorithmic output is *not* based only upon facts that relate to the past actions of the decision-subject, we

[84] Ibid, 71.

https://doi.org/10.1017/9781009334297.016 Published online by Cambridge University Press

cannot justify using it to make decisions. If we do so, we deny the decision-subject the opportunity to avoid the burden by choosing appropriately.

Those who have evaluated COMPAS have challenged both its overall predictive success, and its distribution of the risk of error.[85] But there is an additional problem: each of the COMPAS assessments, most notably the wider 'criminogenic need' assessment, takes into account a range of facts that either have nothing to do with the defendant's actions (such as family background), or which are linked to actions that the defendant could never reasonably have suspected would result in criminal punishment (such as choice of friends or 'associates'). Thus, they deny the defendant a meaningful opportunity to choose to act in a manner that will avoid the risk of criminal punishment. And if the prediction takes into account facts that we have good reason to exclude from the decision, the solution is not to give the predictive output *less* weight (by applying human discretion). It is to give it no weight at all.

11.8 SAFEGUARDS

We cannot safeguard effectively against unjust decisions by applying human discretion to a predictive output at the time of decision-making. Appropriate 'safeguarding' means ensuring that the decision-making tools that we use take into account the right information in the right way, long before they enter our decision-making fora. I have made some concrete recommendations about how to determine whether the ADM/AI tool meets that threshold, which I summarise here.

The first question we should ask is this: is the prediction about what the decision-subject will do? If the answer to that question is no, we can in principle justify using the ADM/AI tool. Whether we can in practice turns on its predictive success – its overall success rate, and how the risk of error is distributed. We can assess these things statistically – without 'opening the black box', and without identifying reasons for any given prediction. If the ADM/AI tool fares just as well or better than humans, we can use it, and we can offer explanations to the decision-subject that are based on how we use it. If it does not fare just as well or better than humans, we cannot.

If the prediction is about what the decision-subject will do, we need to know the reasons for the prediction, which we can determine by using the counterfactual technique. We can only justify using the ADM/AI tool if three conditions are satisfied: (i) as above, the prediction is accurate and the risk of error is distributed evenly; (ii) the prediction is based solely on what the decision-subject has done; and (iii) the defendant has had sufficient opportunity to discover that those actions could result in these consequences.

[85] See e.g. Tim Brennan, William Dieterich, and Beate Ehret, 'Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System' (2009) 36 *Criminal Justice and Behavior* 21.

https://doi.org/10.1017/9781009334297.016 Published online by Cambridge University Press

It bears emphasis that the concern about policies that deny individuals a meaningful opportunity to avoid incurring certain burdens is not confined to the sphere of ADM. Courts in Wisconsin are permitted to take into account educational background and PSI results in sentencing decisions,[86] and the Wisconsin DOC directs agents completing the PSI to take into account a range of factors that include: intelligence; physical health and appearance; hygiene and nutrition; use of social security benefits or other public financial assistance; the nature of their peer group; and common interests with gang-affiliated members.[87] Thus, safeguarding efforts should not merely be directed towards ADM; they should take into account the broader law and policy landscape, of which ADM forms one part.

When we impose burdens on some people for the sake of some benefit to others, we should (wherever possible) present these people with valuable opportunities to avoid those burdens by choosing appropriately. And when the burdens that we impose are as exceptional as criminal incarceration, this requirement is all the more urgent: we cannot justify sending people to prison because they received poor grades in school, because their parents separated when they were young, or because of choices that their friends or family have made; we must base our decision on the choices that they have made, given a range of meaningful alternatives.

[86] *State v Harris*, 119 Wis2d 612, 623, 350 NW 2d 633 (1984).
[87] State of Wisconsin Department of Corrections, *Wisconsin Department of Corrections Electronic Case Reference Manual*.