European Journal on Criminal Policy and Research (2022) 28:407–423
https://doi.org/10.1007/s10610-022-09516-8

Predicting Recidivism Risk Meets AI Act

Gijs van Dijck[1]

Accepted: 31 May 2022 / Published online: 10 June 2022

Abstract
Quantitative recidivism risk assessment can be used at the pretrial detention, trial, sentencing, and / or parole stage in the justice system. It has been criticized for what is measured, whether the predictions are more accurate than those made by humans, whether it creates or increases inequality and discrimination, and whether it compromises or violates other aspects of fairness. This criticism becomes even more topical with the arrival of the Artificial Intelligence (AI) Act. This article identifies and applies the relevant rules of the proposed AI Act in relation to quantitative recidivism risk assessment. It does so by focusing on the proposed rules for the quality of the data and the models used, on biases, and on the human oversight. It is concluded that legislation may consider requiring providers of high-risk AI systems to demonstrate that their solution performs significantly better than risk assessments based on simple models, and better than human assessment. Furthermore, there is no single answer to evaluate the performance of quantitative recidivism risk assessment tools that are or may be deployed in practice. Finally, three approaches of human oversight are discussed to correct for the negative effects of quantitative risk assessment: the optional, benchmark, and feedback approach.

Keywords Recidivism · Quantitative risk assessment · Pre-trial detention · COMPAS · OxRec

Introduction

Predicting recidivism risk is an important activity at the pretrial detention, trial, sentencing, and parole stage. Several scientific models have become available in the past decades that carry out quantitative risk assessments (Harcourt, 2015; McGuire, 2004). Criticism predominantly concerns what the models that are used measure and intend to measure, whether the predictions are more accurate than when humans conduct the risk assessment, and whether the prediction models cause or increase inequality and discrimination or otherwise compromise fairness (e.g., Dressel & Farid, 2018; Kehl et al., 2017; Skeem & Lowenkamp, 2016; Starr, 2014; Završnik, 2019).

✉ Gijs van Dijck
gijs.vandijck@maastrichtuniversity.nl

[1] Professor of Law and Director of the Maastricht Law and Tech Lab at Maastricht University, Maastricht, the Netherlands

The European Commission has fairly recently published a proposal for regulating artificial intelligence (AI). The proposed act, which is titled 'Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act),' carries the abbreviated name 'AI Act,' and will be further discussed later on, is expected to have a significant impact on both the users and providers of quantitative risk assessment tools or software in Europe, and possibly beyond. Even though the AI Act finds itself in the proposal stage, it already becomes clear that it, even in an altered form, will have a significant impact on quantitative recidivism risk assessment. This article explores the circumstances under which quantitative recidivism risk assessments are expected to be compliant with the proposed AI Act. This will be illustrated by means of two risk assessments tools used in practice to predict recidivism.

After a brief introduction of the recidivism risk assessment tools (*Sect. 2*), this article continues with discussing the proposed AI Act. First, the background of the proposed act is discussed, along with the applicability of it to quantitative recidivism risk assessment (*Sect. 3*). Second, the proposed rules on data governance and measurement practices are discussed (*Sect. 4*). This section includes an in-depth examination of the performance of the two risk assessment methods, where the statistical performance is, among other things, compared to human assessment as well as to simple models. The article subsequently continues with discussing the delicate topic of addressing possible biases (*Sect. 5*). In a broad sense, bias is inherent to quantitative recidivism risk assessment. It can be detrimental to suspects or offenders because it can be at odds with the principle of open justice (i.e., denial of oversight of algorithmic tools used for determining a defendant's legal status), equality of arms, equality before courts, the presumption of innocence, the principle of individualized justice, or the right to a fair and public hearing (e.g., McKay, 2020, providing further references). Quantitative risk assessment has in fact been criticized for the lack of testability and contestability, for using aggregate group data to assess individuals, and for the use of undisclosed proprietary information (McKay, 2020). In this article, I focus on bias in a more narrow sense, more specifically on systems that make predictions that results in the selection or preference of a certain outcome as the result of systematic error related to sampling, variable selection, or testing, which causes the predictions to be systematically too high or too low for certain subgroups. Finally, human oversight in relation to recidivism risk assessment is discussed (*Sect. 6*). Concluding remarks complete this contribution (*Sect. 7*).

Recidivism Risk Assessment Tools

Quantitative risk assessment is commonly based on machine learning or statistics. Machine learning concerns the process where a model is built based on sample data, commonly referred to as training data, in order to make predictions or decisions. The difference with a statistical approach is that the machine understands patterns and trains algorithms[1] by itself, whereas statistical approaches rely on mathematical concepts for finding patterns in data as defined by the researcher. In this contribution, I illustrate each approach with one real-life application. I choose *Oxford Risk of Recidivism Tool* (OxRec) as an example

[1] Algorithms can be described as procedures for solving mathematical problems (…) 'in a finite number of steps that frequently involves repetition of an operation,' see Merriam-Webster Online Dictionary (last accessed 22 March 2022).

of a statistical approach, and *Correctional Offender Management Profiling for Alternative Sanctions* (COMPAS) as an example of a machine learning approach.

OxRec is a tool that assists in the prediction of violent reoffending in released prisoners. It is based on scientific research that was conducted among a sample of convicted persons. The initial study was conducted among Swedish convicts (Fazel et al., 2017). Later, the study was validated in the Dutch context (Fazel et al., 2019). It followed the Swedish study as closely as possible in terms of the variables collected and definitions used.[2] The research has led to online tools that can be used to estimate recidivism risk for both Sweden[3] and the Netherlands.[4]

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), in turn, is software, owned by *Equivant* (previously *Northpointe*), that has been used by some US courts to assess recidivism risk. COMPAS is controversial for the reasons mentioned in the introduction (Angelino et al., 2018; Brennan et al., 2008; Mayson, 2019; Rudin et al., 2020). Because the software is proprietary, the data and algorithms are not transparent, also not for the suspect and the judge. This lack of transparency was addressed in *Loomis v. Wisconsin*.[5] In this case, the suspect challenged the quality of the prediction tool and argued that he had the right to be sentenced based on accurate information and that not having insight into how COMPAS comes to its prediction score violates this right. The Wisconsin Supreme Court rejected the suspect's arguments. It argued that, if applied properly, COMPAS can 'enhance a judge's evaluation, weighing, and application of the other sentencing evidence in the formulation of an individualized sentencing program appropriate for each defendant.'

This article discusses OxRec and COMPAS, but the observations equally apply to other actuarial risk assessment tools, such as the Level of Service/Case Management Inventory (LS/CMI) (Andrews et al., 2004, estimating general recidivism risk in adults), the Youth Level of Service/Case Management Inventory (YLS/CMI) (Hoge & Andrews, 2006, the youth version of the LS/CMI), the STATIC-99 (Hanson & Thornton, 2000, estimating sexual recidivism risk in adults), the Structured Assessment of Violence Risk in Youth (SAVRY) (Borum et al., 2006, estimating violent recidivism in juveniles), and the Historical, Clinical and Risk Management – 20 (HCR-20) Version 2 (Webster et al., 1997) and Version 3 (Douglas et al., 2013) (estimating violent recidivism risk in adults). These tools have been developed based on statistical models that, similarly to OxRec and as will be explained below, are likely to be considered AI systems.

[2] See <https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-018-37539-x/MediaObjects/41598_2018_ 37539_MOESM1_ESM.pdf> for an overview and definitions of the variables used and to what extent they agree/diverge from the Swedish study (last accessed 20 June 2020; the url is no longer accessible).

[3] <https://oxrisk.com/krimrec/> (last accessed 17 January 2022).

[4] <https://oxrisk.com/oxrec-nl-2-backup/> (last accessed 17 January 2022).

[5] *Loomis v. Wisconsin*, 881 N.W.2d 749 (Wis. 2016). See already *State v. Samsa*, 2015 WI App 6, 359 Wis.2d 580, 859 N.W.2d 149 (court of appeals approving a circuit court's consideration of a COMPAS assessment at sentencing, while stating 'COMPAS is merely one tool available to a court at the time of sentencing' (at 359).

Artificial Intelligence Act (AI Act)

On 21 April 2021, the European Commission published a proposed regulation for artificial intelligence. It follows from a White Paper on AI,[6] from calls of the European Council for identifying high-risk AI systems in relation to fundamental rights and enforcement of legal rules,[7] and from the 2020 and 2021 adoption of AI-related resolutions on ethics,[8] liability,[9] copyright,[10] criminal matters[11] and education, culture and the audio-visual sector.[12] The proposed AI Act aims to 'improve the functioning of the internal market by laying down a uniform legal framework in particular for the development, marketing and use of artificial intelligence in conformity with Union values' (Recital 1). Under the AI Act, an AI system is defined as:

> 'software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with' (Article 3(1)).

The approaches listed in Annex I include machine learning approaches (supervised and unsupervised), logic-based and knowledge-based approaches (e.g., expert systems), and statistical approaches (e.g., Bayesian estimation, search and optimization methods). The AI Act applies to providers and users of AI systems that are placed on the market or used in the Union, or of systems of which the output is used in the Union, irrespective of where the provider is established (Article 2). Users[13] such as the parole officer (and organization) and judges (courts), as well as providers[14] such as the software creator come to mind as those who are subjected to the rules laid down in the AI Act. Considering that quantitative recidivism risk assessments commonly come in the form of a set of instructions used to execute specific tasks by a computer, instruments or tools are likely to be considered software.

AI systems are considered 'high-risk' if they fulfill the criteria of Article 6(1) or are listed in Annex III of the AI Act. Article 6(1) requires the AI systems to be a 'safety

[6] European Commission, White Paper on Artificial Intelligence—A European approach to excellence and trust, COM(2020) 65 final, 2020.

[7] European Council, Special meeting of the European Council (1 and 2 October 2020) – Conclusions EUCO 13/20, 2020; Council of the European Union, Presidency conclusions—The Charter of Fundamental Rights in the context of Artificial Intelligence and Digital Change, 11,481/20, 2020.

[8] European Parliament resolution of 20 October 2020 on a framework of ethical aspects of artificial intelligence, robotics and related technologies, 2020/2012(INL).

[9] European Parliament resolution of 20 October 2020 on a civil liability regime for artificial intelligence, 2020/2014(INL).

[10] European Parliament resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies, 2020/2015(INI).

[11] European Parliament Draft Report, Artificial intelligence in criminal law and its use by the police and judicial authorities in criminal matters, 2020/2016(INI).

[12] European Parliament Draft Report, Artificial intelligence in education, culture and the audiovisual sector, 2020/2017(INI). In that regard, the Commission has adopted the Digital Education Action Plan 2021–2027: Resetting education and training for the digital age, which foresees the development of ethical guidelines in AI and Data usage in education – Commission Communication COM(2020) 624 final.

[13] '[A]ny natural or legal person, public authority, agency or other body using an AI system under its authority' (Article 3(4) AI Act).

[14] '[N]atural or legal person, public authority, agency or other body that develops an AI system or that has an AI system developed with a view to placing it on the market or putting it into service under its own name or trademark, whether for payment or free of charge' (Article 3(2) AI Act).

component of a product, or [that] itself [is] a product' and a 'product whose safety component is the AI system, or the AI system itself as a product, is required to undergo a third-party conformity assessment with a view to the placing on the market or putting into service.' For the purpose of this article, the examples mentioned in Annex III are of relevance. This particularly applies to number 6 and number 8 of Annex III. Number 6 stipulates when AI systems used by law enforcements are considered high-risk:

> "(a) AI systems intended to be used by law enforcement authorities for making individual risk assessments of natural persons in order to assess the risk of a natural person for offending or reoffending or the risk for potential victims of criminal offences;
> (…)
> (e) AI systems intended to be used by law enforcement authorities for predicting the occurrence or reoccurrence of an actual or potential criminal offence based on profiling of natural persons as referred to in Article 3(4) of Directive (EU) 2016/680 or assessing personality traits and characteristics or past criminal behaviour of natural persons or groups;
> (f) AI systems intended to be used by law enforcement authorities for profiling of natural persons as referred to in Article 3(4) of Directive (EU) 2016/680 in the course of detection, investigation or prosecution of criminal offences;
> (…)".

One can debate under which of the categories quantitative recidivism risk assessment falls. For instance, it is debatable whether it can be qualified as profiling, which is defined in Article 3(4) of Directive 2016/680[15] as:

> 'any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements'.

Both COMPAS and OxRec fall under this definition, as the tools capture and use 'certain personal aspects relating to a natural person' such as income, marital status, prior alcohol abuse, drug use, and psychological illness of the suspect or alleged offender. In any case, recidivism risk predictions fall under 'individual risk assessments of natural persons in order to assess the risk of a natural person for offending or reoffending or the risk for potential victims of criminal offences.'

Number 8 of Annex III, in turn, explains when an AI system is considered high-risk when applied in the context of 'Administration of justice and democratic processes' (Schwemer et al., 2021):

> (a) AI systems intended to assist a judicial authority in researching and interpreting facts and the law and in applying the law to a concrete set of facts.

Recital 40 provides some clarification on number 8 of the Annex. It explains that:

[15] Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA.

> '[c]ertain AI systems intended for the administration of justice and democratic processes should be classified as high-risk, considering their potentially significant impact on democracy, rule of law, individual freedoms as well as the right to an effective remedy and to a fair trial. In particular, to address the risks of potential biases, errors and opacity, it is appropriate to qualify as high-risk AI systems intended to assist judicial authorities in researching and interpreting facts and the law and in applying the law to a concrete set of facts. Such qualification should not extend, however, to AI systems intended for purely ancillary administrative activities that do not affect the actual administration of justice in individual cases, such as anonymisation or pseudonymisation of judicial decisions, documents or data, communication between personnel, administrative tasks or allocation of resources'.

In the context of judicial adjudication, the main question is whether the active (courts using quantitative risk assessments to assess recidivism risk) or passive (considering or accepting quantitative recidivism risk assessments presented by one of the parties) use of quantitative recidivism risk assessments will fall under 'intended to assist judicial authorities in researching and interpreting facts and the law and in applying the law to a concrete set of facts.' I would argue it would, considering that a risk assessment can be considered a 'fact' that the law applies to and that can be interpreted and weighed against the other facts of the case.

Article 9 and following lay down the rules that apply to high-risk AI systems. Article 9 requires a risk management system that includes a risk assessment of the known and foreseeable risks when the AI system is used in accordance with 'its intended purpose and under conditions of reasonably foreseeable misuse.' In addition, it requires an evaluation of possible risks based on the so-called 'post-market monitoring system,' which is defined as activities of the AI provider to collect and review experience from the AI system. Finally, Article 9 requires the adaptation of risk management measures, which includes the testing of the AI system

> 'against preliminarily defined metrics and probabilistic thresholds that are appropriate to the intended purpose of the high-risk AI system' (Article 9(7) AI Act).

Users of high-risk AI systems such as parole boards and courts should, among other things, use and monitor such systems in accordance with the instructions of use, and ensure that the user input is relevant in view of the intended purpose (Article 29(1–5) AI Act). Providers, in turn, are obliged to, essentially, comply with the registration obligations of Article 51, affix the CE marking to a high-risk AI systems to indicate the conformity with the Regulation, comply with rules laid down in the AI Act, take corrective actions if the AI system does not comply with the AI Act, inform the national competent authorities of non-compliance and corrective actions, and demonstrate conformity upon request of a national competent authority (Article 16).

Market surveillance authorities 'shall be granted full access to the training, validation and testing datasets used by the provider' (Article 64 AI Act). This is a potentially far-reaching provision, as it would allow such authorities to have access to and evaluate the data, and to test algorithms that were developed based on the data, even in case of proprietary software. If the software or tooling presents a risk to the health or safety of persons, to the compliance with fundamental rights, or to other aspects of public interest protection, the market surveillance authority must take appropriate measures, which may include withdrawing the AI system from the market, even if the AI system is otherwise compliant with the AI Act (Article 67(1) AI Act). Depending on the violation, non-compliance with

obligations in the AI Act can lead to administrative fines of a maximum between EUR 500,000 and EUR 30,000,000, or, in case of a company, up to 2%, 4%, or 6% of its total worldwide annual turnover (Article 72 AI Act).

Data Governance and Measurement Practices

The proposed AI Act aims to ensure that the data and models that are used for the AI system are of a certain quality. In this respect, Article 10 stipulates that AI systems that deploy training, validation, and testing datasets 'shall be subject to appropriate data governance and management practices' (Article 10(2) AI Act). COMPAS' software clearly falls under this definition, as its algorithm is based on machine learning models where the data are split in a training set on which the algorithm is trained, and in a test set on which the performance of the algorithm is evaluated.

For OxRec (and OxRec-like tooling), it is more debatable whether the software falls under the definition of Article 10(2) AI Act. Unlike COMPAS, OxRec is based on a statistical model where it is transparent how the variables contribute to the prediction of the outcome variable. This qualifies OxRec as an AI system under the definition of Article 3(1) and Annex I, yet not necessarily as one that falls under Article 10(1) AI Act. However, a closer inspection, and as will be further illustrated below when discussing the results of OxRec, reveals that OxRec has deployed a strategy of using training[16] and test[17] sets to calculate model performance. From this, I would be inclined to conclude that also OxRec is subject to appropriate data governance and management practices.

Data governance and management practices concern design choices, data collection, data preparation processes, the formulation of assumptions, assessments of the availability, quantity, and suitability of the data, bias examination, and the identification of, and solutions for, 'data gaps or shortcomings' (Article 10(2) AI Act). Furthermore, data governance and management practices require the datasets to be relevant, representative, free of errors, and complete. They must have the appropriate statistical properties (Article 10(3) AI Act) and must take into account 'the characteristics or elements that are particular to the specific geographical, behavioral or functional setting within which the high-risk AI system is intended to be used' (Article 10(4) AI Act). A lot can be said about each individual data governance and management practice. For the purpose of this article, and for reasons of feasibility, I focus on the suitability of the data and the models deployed.

The performance of the two prediction models can be measured by means of a number of commonly used metrics. Both OxRec and COMPAS report measures such as sensitivity, specificity, positive predicted values, negative predicted values, and Area Under Curve (AUC). These evaluation metrics essentially rely on true and false positives and on true and false negatives:

[16] '[M]eans data used for training an AI system through fitting its learnable parameters, including the weights of a neural network' (Article 3(29) AI Act).

[17] '[D]ata used for providing an independent evaluation of the trained and validated AI system in order to confirm the expected performance of that system before its placing on the market or putting into service' (Article 3(31) AI Act).

- Sensitivity: the ability of an instrument to correctly identifying individuals belonging to a particular risk group.[18]
- Specificity: the ability of an instrument to correctly identifying individuals *not* belonging to a particular risk group.[19]
- Positive predicted values: the probability that the instrument triggers a false alarm.[20]
- Negative predicted values: the probability that the instrument remains silent where the alarm should have been triggered.[21]
- Area Under Curve (AUC): uses true and false positives and negatives to quantify the adequacy of a prediction instrument by distinguishing individuals in a particular risk group from individuals who are not in that risk group. An AUC of 1.0 does this perfectly (no false alarms, no false negatives), whereas a score of 0.5 indicates that the models performs no better than a coin flip. What is considered an acceptable AUC value, depends on the classification task.

These metrics can be used to determine whether it is preferred that those who belong to a certain recidivism risk category (e.g., 'high') is classified as such, or whether it should be prevented that individuals end up being classified in the wrong category (e.g., 'high risk' when the risk is not high). Risk thresholds can be used to find a balance between false positives and false negatives. A low threshold will lead to the correct identification of those who belong to a certain risk group, but also to the designation of persons who do not belong to it (false alarms). Conversely, a high threshold ensures a small probability of persons being placed in a risk group to which they do not belong, but it also results in a larger number of individuals who belong to a certain risk group yet go unnoticed.

The Dutch OxRec study reveals c-scores, which are comparable to AUC scores, of between 0.67 and 0.69, with 95% certainty that the true scores lie between 0.65 and 0.70 (Fazel et al., 2017, 2019). The Swedish study reported a c-index of 0.74 (AUC of 0.76 over two years) (Fazel et al., 2016a). The reported AUC for COMPAS reportedly lies between 0.68 and 0.73 (Flores et al., 2016). These results suggest that the probability of a correct classification is better than the flip of a coin (AUC = 0.50), but that there is still a significant probability of wrong classifications.

Whether AUC scores, or other scores, are acceptable, depends on the available alternatives. A vaccine that is 30% effective might still be considered acceptable if the alternative is 0% effectiveness. This already shows the difficulty of assessing the data governance and management practices referred to in the AI Act: there are no straightforward rules or guidelines on when certain practices fall below or above 'the' threshold, which in itself can be an arbitrary value. In that way, quantitative risk assessment to an important extent relies on a risk-based assessment made by one or more individuals prior to its application.

In the situation of quantitative risk assessment, it would be logical to compare its results with how humans perform when predicting recidivism risk. The evidence of the performance of statistical methods, compared to human prediction, is mixed. Research on clinical prediction has found that statistical methods frequently outperform clinical assessments (Wolff, 2008; Starr, 2014; Oleson, 2011, at 1342(fn84); Gottfredson & Moriarty, 2006; Harris, 2006; Ægisdóttir et al., 2006). The number of studies that compare human

[18] True positive rate = true positives / (true positives + false negatives).
[19] True negative rate = true negatives / (true negatives + false positives).
[20] PPV = true positives / (true positives + false positives).
[21] NPV = true negatives / (true negatives + false negatives).

judgments and statistical prediction in the area of predicting recidivism risk are, however, limited. Dressel and Farid conducted an experiment where they compared the assessments produced by COMPAS to those of individuals with little or no expertise in the field of criminal justice (Dressel & Farid, 2018). A comparison of the results revealed similar accuracy scores: 62.8% accuracy (AUC-ROC = 0.71) for participants versus 65.2% (AUC-ROC = 0.70) for COMPAS. When pooling participant responses (20 responses per subset) and applying a majority rules criterion within each subset, a crowd accuracy of 67.0% was obtained, higher than COMPAS yet not statistically significant (p = .085). Both predictions, however, may have been overestimations. The accuracy for COMPAS' actual recidivism risk might be lower than reported, considering that the algorithm might partly be predicting policing bias (i.e., the police targeting specific groups with high recidivism risks). The participants, in turn, received feedback on whether their prediction was correct and on their average accuracy level after each prediction.

In a replication study, Lin et al. (2020) found almost identical results. COMPAS' predictions were made with 65% accuracy, against 64% for human predictions. Even without participants receiving immediate feedback, the accuracy was still 62%. However, in experiments that extended the initial study to non-COMPAS instruments, the researchers did find that algorithms outperformed humans when predicting recidivism risk. Differences between algorithmic and human predictions were particularly observed when participants were not provided immediate feedback on the accuracy after a prediction (compared to receiving feedback). Algorithmic and logistic regression predictions resulted in 89% accuracy, against 83% for human prediction with feedback after each prediction, and 60% without feedback. Little to no improvement in classification accuracy was found when participants were provided information about ten more risk factors on which to base predictions. Human predictions were on par with statistical predictions with a low number of risk factors (n = 5), whereas statistical predictions outperformed human predictions when the number of risk factors was high (n = 15).

The number of variables matters when comparing predictions of humans and machines, but it remains questionable whether models with a large number of variables are necessary in order to make accurate predictions; whether one cannot rely on a simply model with only a few variables. The first evidence that simple algorithmic rules can have a strong predictive power and outperform human clinical assessments can be traced back to at least 1955, when researchers compared parole predictions of four sociologists and four psychiatrists to predictions made by a statistical model with seven predictors (Glaser, 1955). Statistical models and artificial intelligence techniques, machine learning in particular, have since gained popularity, making it common to analyze large datasets with a large number of predictors (like COMPAS: 137 predictors) to make estimates of certain outcomes. Interestingly, prediction models that use sophisticated artificial intelligence and a large number of predictors do not necessarily offer substantially better predictions than simple models, particularly when the human behavior is predicted, that is predictions of outcomes that involve human thinking or human decision-making. The Dressel and Farid (2018) study already suggested that a simple logistic regression classifier with only seven features yields similar accuracy (66.6%) as does COMPAS, which relies on 137 features (65.4%). Moreover, a classifier with only two features performed as well as COMPAS. The two features, age and total number of previous convictions, followed from two meta-analysis studies as the features with the highest predictive power (66.8%).

More evidence for the power of simple models comes from Jung et al. (2017), who developed a, what they called, 'select-regress-round' approach and compared it to the results produced by sophisticated AI for 22 publicly available datasets from a machine

learning repository.[22] The select-regress-round approach consists of three steps. First, a number of k features is selected from the total number of features available (*select*). Second, a logistic regression model is trained on the selected features (*regress*). Third, the coefficients are rescaled in a way that the results become more intuitive to individuals with little or no statistical background. The coefficients are subsequently rounded to the nearest integer [-M,M] (*round*). Remarkably, the select-regress-round approach with a maximum of five features and rounding coefficients between -3 and 3 yielded a mean AUC score of 87%, close to the mean 92% AUC score of a random forest model with between 11 and 93 features (38 on average).

A more recent study compared the findings of 160 (!) research teams to the results of a single domain expert (Salganik et al., 2020). The task was to predict life outcomes, something sociologists had been investigating for a substantive amount of time. The participating research teams had access to a large dataset with thousands of variables and were allowed and encouraged to use the techniques that they deemed most fit, including the most modern and complex machine learning models. The findings of the 160 research teams were compared to the results of one domain expert who used a simple logistic or linear regression classifier trained on a handful of variables: race / ethnicity, marital status, and the mother's educational level at child's birth. The comparison revealed that the prediction scores were generally low overall, regardless of model that was trained or chosen. The association between the prediction error and the prediction technique was weak. Consequently, the results indicate that the scientific mass collaboration only slightly outperformed logistic or linear regression models with a handful of features selected by the domain expert.

The examples suggest that an increased number of features and sophisticated, sometimes resource-draining models, do not necessarily substantively improve predictions. A reason for why simple heuristics perform fairly well compared to sophisticated models, and why the predictive power of recidivism risk predictions is limited, may lie in what is predicted: human behavior. Different individuals tend to make different predictions based on the same input, which is likely the result of anchoring effects (i.e., making estimates based on an initial value or position – the anchor) (Tversky & Kahneman, 1974), order effects (i.e., the order in which information is presented) (Blankenship, 1942), and other cognitive biases and human decision-making processes that are likely to impact human judgments and their consistency.

These insights beg the question what can reasonably be expected from users of quantitative recidivism risk assessment software as to the evaluation of that software. What conclusion should a user draw if a model with a few predictors, or a select-regress-round approach performs similarly to complex black box (such as COMPAS) or even a non-black-box yet (such as OxRec) tooling? Should s/he actively search for evidence regarding the use of such tools? And how to weigh competing insights and evidence? It seems a lot to ask from, for instance, parole officers, prosecutors, or judges. Providers should be better equipped to make comparisons with alternatives that may exist or could be created. However, providers have an interest in advocating for their tooling or software, especially if there is a business model attached to it. Consequently, the legislator may consider including a rule in the final version of the AI Act that stipulates that providers of high-risk AI systems should provide evidence of that their solution performs significantly better than risk assessments based on simple models, and better than human assessment.

[22] <https://archive.ics.uci.edu/ml/index.php> (last accessed 17 January 2022).

Possible (Fairness) Biases

Article 10(2)(f) of the proposed AI Act stipulates that data governance and management practices shall include an 'examination in view of possible biases.' As pointed out in the introduction, the possibility that quantitative recidivism risk assessment can be discriminatory or otherwise cause or exacerbate inequality is well documented (e.g., Dressel & Farid, 2018; Kehl et al., 2017; Skeem & Lowenkamp, 2016; Starr, 2014; Završnik, 2019). Considering that this (and other) previous research has extensively documented the potentially biased character of COMPAS, I will here mostly focus on OxRec.

The study conducted in the Netherlands used most of the variables and definitions that were also used in the Swedish study.[23] The research included a range of predictors, including gender, age, length of incarceration, relationship status (single/other), level of education, income, alcohol use, drug use, and mental illness.[24] Particularly the *Deprivation* variable is interesting from the perspective of ethnic profiling. This variable consists of a number of indicators, namely postal code, welfare recipiency, unemployment, low level of education, crime rate, and median income.

Several of these variables and indicators may be classified as potentially discriminatory or causes of inequality. Postal code, gender, age, education level, and income can be linked to recidivism, but they can also be (partly) direct or indirect indicators of ethnicity or social classes. If persons in certain population groups, with a certain ethnicity, in a certain social class, etc., are more likely to be arrested and convicted than others, the aforementioned variables will reflect this information (Frase, 2009). A prediction model can recognize discriminatory patterns, which will (further) disadvantage certain groups. This becomes apparent when exploring the online tool (Braverman et al., 2016).[25]

Possible inequality as a result of use of the software can be the result of the algorithm or statistical procedure, or of the data that served as input for training the statistical or machine learning models. One solution is to not include questionable variables such as neighborhood deprivation. However, not including variables can result in prediction models that are more discriminatory than a model with the variables (e.g., higher risk estimation for women when gender variable is omitted) (Monahan & Skeem, 2016). Moreover, the problem is that many historical variables can be potentially discriminatory (Fazel et al, 2016b).

Another solution for ruling out profiling (e.g., ethnic, social) is to measure possible biases by collecting additional data that captures information about individuals. One may, for instance, collect information about the ethnic or social background of the individuals involved in the study and analyze whether the application of the risk assessment models or software results in different outcomes for certain ethnic or social groups. Article 10(5) of the proposed AI Act would allow for such data collection and analysis, since it states that

> '[t]o the extent that it is strictly necessary for the purposes of ensuring bias monitoring, detection and correction in relation to the high-risk AI systems, the providers of

[23] An overview is available at <https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-018-37539-x/MediaObjects/41598_2018_37539_MOESM1_ESM.pdf> (last accessed 17 January 2022).

[24] In the Swedish study, immigrant status was still included as a variable, but this information was not available in the datasets used for the Dutch study.

[25] <https://oxrisk.com/oxrec/> (Sweden), <https://oxrisk.com/oxrec-nl-2-backup/> (the Netherlands) (last accessed 17 January 2022).

such systems may process special categories of personal data referred to in Article 9(1) of Regulation (EU) 2016/679, Article 10 of Directive (EU) 2016/680 and Article 10(1) of Regulation (EU) 2018/1725'.

Understandably, carrying out such data collection or analysis is subject to appropriate safeguards, which includes technical limitations on the re-use and use of privacy-preserving strategies like pseudonymization or encryption.

More or better data, or improved models, however, will not be able to solve the fundamental underlying issue. The problem with quantitative risk assessment is that fairness, whether it concerns discrimination or other possible differences between certain groups, can be operationalized in two ways: (1) by keeping the error rates comparable between groups, or (2) by treating those with the same risk scores in the same way. As excellently explained in Hao & Stray, 2019, both operationalizations are defensible, but satisfying them at the same time is impossible. If a certain ethnic group is more likely to be rearrested compared to another ethnic group, the percentage of wrongly classified individuals when predicting recidivism will also differ between the two groups. One can correct for this by changing the release and detention threshold for the two groups (i.e., the threshold for the recidivism risk to be sufficiently high to detain the person), making the threshold higher for one group so that the proportion of detained persons (compared to released persons) is similar for the two groups. This ensures that there is no discrimination between the two groups: regardless of the ethnic group one belongs to, the probability of being detained is the same (all other things being equal). However, the result of this correction is that of the two persons who receive the same risk score and belong to different ethnic groups, one will be detained whereas the other will be released. In other words, two persons with the same risk score do not receive the same treatment. This sounds discriminatory, but correcting for this comes down to setting the same risk thresholds for both groups, which results in returning to the original situation, with members of a certain ethnic group more likely to be rearrested compared to members of another ethnic group. The problem lies in the data that was used for developing models for the quantitative risk assessment: if persons from different groups are rearrested at different rates, it leads to a problem that cannot be solved with statistics or with sophisticated machine learning approaches such as deep learning and neural networks.

Considering the complexity of handling possible discrimination bias in quantitative risk assessment, how can an examination in view of possible biases as referred to in Article 10(2)(f) of the proposed AI Act take shape? In their article, Eckhouse et al. (2019) propose a layered approach for assessing the fairness of quantitative risk assessment. Their framework consists of three questions:

1. Is it fair to use data about other people to make decisions about an individual?
2. Are the data used biased in a fundamental way?
3. Is the model that assigns a risk score fair?

The framework is layered in that each layer depends on the previous layer. Data quality or model performance does not matter if making judgments about individuals based on groups is considered unfair or illegitimate, and a fair model will produce biased outcomes if the data are biased. The layered approach can be used by organizations and individuals to at least make transparent which risks of biases in the risk assessment were considered and which of those risks were deemed acceptable. All three layers need to be fair for criminal justice risk assessments.

Human Oversight

The AI Act, if it were to be enacted in its current version, requires human oversight for high-risk AI systems. Article 14 (and Recital 48) of the proposed act stipulates that high-risk AI systems must be designed in such a way that it can be overseen by humans when the system is in use. Section 4 of Article 14 provides a list of the purposes that the measures to achieve human oversight should have. Those who human oversight is assigned to should:

> '(a) fully understand the capacities and limitations of the high-risk AI system and be able to duly monitor its operation, so that signs of anomalies, dysfunctions and unexpected performance can be detected and addressed as soon as possible;
> (b) remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system ('automation bias'), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons;
> (c) be able to correctly interpret the high-risk AI system's output, taking into account in particular the characteristics of the system and the interpretation tools and methods available;
> (d) be able to decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the high-risk AI system;
> (e) be able to intervene on the operation of the high-risk AI system or interrupt the system through a "stop" button or a similar procedure'.

Article 14 AI Act sets high standards for 'those human oversight is assigned to.' The purpose, to have this person 'fully understand' the possibilities and limitations of the AI system, requires in-depth knowledge of the data, the statistical methods, and / or the algorithm on which the AI system is based, information the person will not always have access to. The same applies to the purpose of 'correctly interpret[ing]' the output of the AI system, as there is no such thing as 'the' correct interpretation of an AI system. What is 'correct,' depends on what the purpose of the system is, the quality of the data, the design choices of the method that was deployed, and what error levels one is willing to accept. Consequently, a variety of expertise needs to be gathered in order to properly assess the standards in Article 14. How human oversight is organized will therefore be the decisive factor for proper human oversight.

It seems reasonable to allow for deviating from the output of the AI system, yet it raises the question where and how exactly the human should be in the loop. For instance, a situation where certain users within the same organization frequently put the AI system aside and others usually rely on the system seems a rather arbitrary operationalization of the requirement. At least three approaches can be distinguished that may offer guidance as to how to combine human assessment with quantitative risk assessment.

A first approach is the *optional approach*. Under this approach, a human decides whether and how to assess and weigh the predicted recidivism risk. S/he may consider the application of quantitative risk assessment tools undesirable, for instance because of special circumstances of the case or of the individual whose recidivism risk is estimated, or s/he may reject the application altogether, for example because of the conviction that the predictions are based on data from a country or period that is not representative for the case or individual at hand. The advantage of this model is that it provides full autonomy and flexibility to the decision-maker, whether it is the parole officer or judge, to consider the

information, and to weigh it against other circumstances that may be relevant. Arbitrariness can also be the consequence. Individuals affected by the predictions may experience differences in assessment (both procedurally as in terms of outcomes) depending on the person of the decision-maker.

A second approach is the *benchmark* approach. Under this approach, the quantitative risk assessment is the first step, as it serves as the baseline that the human decision-maker can follow or deviate from. For example, the parole officer computes a risk assessment and subsequently makes a determination of whether the recommendation should be followed. If unregulated or not governed, the benchmark approach produces arbitrariness, as it can leave lots of discretion to the human with respect to whether and when can or should be deviated by the recommendation. Guidelines may be drafted to provide clarity and to increase consistency across different assessors (e.g., parole officers). Such guidelines may include substantive guidelines (e.g., more severe crimes or consequences can cause one to deviate) or procedural ones (e.g., 'comply or explain,' where deviations are systematically collected, analyzed, and used to inform assessors about patterns in the data). What distinguishes the benchmark approach from the optional approach is that quantitative risk assessment serves as an anchor in the former yet not necessarily in the latter. Anchoring effects that come with such a benchmark, can be viewed either positively or negatively, depending on whether one finds it acceptable that information on the group level is the starting point for decisions on the individual level (Eckhouse et al., 2019) and on whether one believes the (cor)relations between the features and the outcomes of the predictions to be causal, reliable, and valid.

A third approach is the *feedback* approach. Here, assessing recidivism risk is considered an iterative process where a human initially assesses the recidivism risk and subsequently compares the results of the human assessment to the prediction made by quantitative risk assessment. The decision-maker may use the information from the prediction model to reflect on the initial assessment and to possibly adjust the assessment. For instance, the parole officer makes his/her assessment and subsequently compares the outcome with the quantitative risk assessment. The feedback model addresses the downsides of the benchmark model. However, it puts the human at the center of the decision-making process, which may lead to less accurate predictions than if statistical or machine learning predictions are more accurate than human predictions. All three approaches can be compliant with Article 14(4)(d) AI Act, although the benchmark and feedback approach seem less arbitrary than is the optional approach.

Conclusion

This article explored the impact the AI Act is expected to have on quantitative risk assessment deployed in the criminal justice system, whether it is used at the pretrial detention, trial, sentencing, or parole stage. It can be concluded that COMPAS and OxRec are high-risk AI systems that would need to comply with the proposed rules on data governance and measurement practices. The available empirical evidence suggests that statistical decision-making generally outperforms clinical decision-making. In addition, sophisticated statistical and machine learning models with many features only fare somewhat better than models that use traditional analyses with a limited number of features, whereas the transparency, explainability, and understandability are often higher in the simple models. Future legislation may consider requiring providers of high-risk AI systems to demonstrate that

their risk assessment model performs significantly better than those based on simple models, and better than human assessment. From a user perspective, the observations suggest that there is no single answer to evaluate the performance of quantitative recidivism risk assessment tools that are or may be deployed in practice, and that it will be practically impossible for users of quantitative risk assessment tools to evaluate them in the way the AI Act expects them to. This particularly also holds true for identifying and addressing possible biases, given that (1) identifying biases requires the availability of data that is often not available, (2) fairness can be defined in more than one way (and in a conflicting way), which makes it difficult for a user to evaluate the appropriateness of a high-risk AI system if the person evaluating is not offered guidance as to which aspects of fairness should be dominant or decisive, and (3) reducing bias is problematic if the underlying data unjustifiably treat social or ethnic groups differently.

Human oversight allows for correcting for the negative effects of quantitative risk assessment, whether as a result of bias or otherwise. In this respect, three approaches were discussed: the optional, benchmark, and feedback approach. As each of the approaches come with their own advantages and limitations, there is no absolute favorite. Regardless, distinguishing between the optional, benchmark, and feedback model can prove valuable in justifying human oversight in the context of quantitative recidivism risk assessment. However, and importantly, if humans are not better at predicting recidivism than are machines, the idea of human oversight raises the question in which ways human oversight can take place in a meaningful way.

In any case, the proposed AI Act, even if it undergoes changes before it is enacted, will have significant impact on the practice of quantitative recidivism risk prediction. It would make explicit various responsibilities for providers and users. If strictly applied, the rules could lead to significant changes in the justification and use of quantitative risk assessment tooling or software, or even, in the absence of proper justifications or measures, to the abandonment of the use of such tools or software. The challenges of quantitative recidivism risk prediction will therefore remain topical.

Funding Not applicable.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Conflicts of interest/Competing interests Not applicable.

References

Andrews, D. A., Bonta, J., & Wormith, J. (2004). The level of service/case management inventory (LS/CMI) [measurement instrument]. Multi-Health Systems.

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., & Rush, J. D. (2006). The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. *The Counseling Psychologist, 34*(3), 341–382. https://doi.org/10.1177/0011000005285875

Angelino, E., et al. (2018). Learning Certifiably Optimal Rule Lists for Categorical Data. *Journal of Machine Learning Research, 18*(234), 1–78.

Braverman, D. W., Doernberg, S. N., Runge, C. P., & Howard, D. S. (2016). OxRec model for assessing risk of recidivism: Ethics. *The Lancet Psychiatry, 9*, 808–809. https://doi.org/10.1016/S2215-0366(16)30175-4

Blankenship, A. (1942). Psychological Difficulties in Measuring Consumer Preference. *Journal of Marketing, 6*(4, Part 2), 66–75. https://doi.org/10.1177/002224294200600420.1

Borum, R., Bartel, P., & Forth, A. (2006). *Manual for the structured assessment of violence in youth (SAVRY)*. Psychological Assessment Resources.

Brennan, T., Dieterich, W., & Ehret, B. (2008). Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System. *Criminal Justice and Behavior, 36*(1), 21–40. https://doi.org/10.1177/0093854808326545

Douglas, K. S., Hart, S. D., Webster, C. D., & Belfrage, H. (2013). *HCR-20^{V3}: Assessing risk for violence – User guide*. Mental Health, Law, and Policy Institute, Simon Fraser University.

Dressel, J. & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1). https://doi.org/10.1126/sciadv.aao5580

Eckhouse, L., Lum, K., Conti-Cook, C., & Ciccolini, J. (2019). Layers of Bias: A Unified Approach for Understanding Problems With Risk Assessment. *Criminal Justice and Behavior, 46*(2), 185–209. https://doi.org/10.1177/0093854818811379

Fazel, S., Chang, Z., Fanshawe, T., Långström, N., Lichtenstein, P., Larsson, H., & Mallett, S. (2016a). Prediction of violent reoffending on release from prison: Derivation and external validation of a scalable tool. *The Lancet Psychiatry, 3*(6), 535–543. https://doi.org/10.1016/S2215-0366(16)00103-6

Fazel, S., Chang, Z., Långström, N., Fanshawe, T., & Mallett, S. (2016b). OxRec model for assessing risk of recidivism: Ethics - Authors' reply. *The Lancet Psychiatry, 3*(9), 809–810. https://doi.org/10.1016/S2215-0366(16)30216-4

Fazel, S., Wolf, A., Larsson, H., Lichtenstein, P., Mallett, S., & Fanshawe, T. R. (2017). Identification of low risk of violent crime in severe mental illness with a clinical prediction tool (Oxford Mental Illness and Violence tool [OxMIV]): A derivation and validation study. *The Lancet Psychiatry, 4*(6), 461–468. https://doi.org/10.1016/S2215-0366(17)30109-8

Fazel, S., Wolf, A., & Vazquez-Montes, M. D. L. A. (2019). Prediction of violent reoffending in prisoners and individuals on probation: A Dutch validation study (OxRec). *Scientific Reports, 9*(1), 841. https://doi.org/10.1038/s41598-018-37539-x

Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). 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. *Federal Probation, 80*(2), 38–46.

Frase, R. S. (2009). What Explains Persistent Racial Disproportionality in Minnesota's Prison and Jail Populations? *Crime and Justice: A Review of Research*, p. 201–280. https://doi.org/10.1086/599199

Glaser, D. (1955). The Efficacy of Alternative Approaches to Parole Prediction. *American Sociological Review, 20*(3), 283–287.

Gottfredson, S. D. & Moriarty, L. J. (2006). Clinical Versus Actuarial Judgments in Criminal Justice Decisions: Should One Replace the Other? *Federal Probation*, 70(2)

Hanson, R. K., & Thornton, D. (2000). Improving risk assessments for sex offenders: A comparison of three actuarial scales. *Law and Human Behavior*, 24(1), 119–136. https://doi-org.mu.idm.oclc.org/10.1023/A:1005482921333

Hao, K. & Stray, J. (2019). Can you make AI fairer than a judge? Play our courtroom algorithm game. Retrieved January 17, 2022, from https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/

Harcourt, B. E. (2015). Risk as a Proxy for Race: The Dangers of Risk Assessment. *Federal Sentencing Reporter, 27*(4), 237–243. https://doi.org/10.1525/fsr.2015.27.4.237

Harris, P. M. (2006). What Community Supervision Officers Need to Know About Actuarial Risk Assessment and Clinical Judgment. *Federal Probation Journal*, 70(2).

Hoge, R. D., & Andrews, D. A. (2006). Youth level of service/case management inventory (YLS/CMI) user's manual. *Multi-Health Systems*. https://doi.org/10.1037/t05078-000

Jung, J., Concannon, C., Shroff, R., Goel, S. & Goldstein, D.G. (2017). Simple rules for complex decisions. https://arxiv.org/abs/1702.04690

Kehl, D., Guo, P. & Kessler, S. (2017). Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing. Responsive Communities Initiative, Berkman Klein Center for Internet & Society, Harvard Law School.

Lin, Z.J., Jung J., Goel, S. & Skeem, J. (2020). The limits of human predictions of recidivism. *Science advances,* 6(7). https://doi.org/10.1126/sciadv.aaz0652

Mayson, S. G. (2019). Bias in, Bias out. *Yale Law Journal, 128*(8), 2218–2301.

McKay, C. (2020). Predicting risk in criminal procedure: Actuarial tools, algorithms, AI and judicial decision-making. *Current Issues in Criminal Justice, 32*(1), 22–39. https://doi.org/10.1080/10345329.2019.1658694

McGuire, J. (2004). Minimising harm in violence risk assessment: Practical solutions to ethical problems? *Health, Risk & Society, 6*(4), 327–345. https://doi.org/10.1080/13698570412331323225

Monahan, J., & Skeem, J. L. (2016). Risk Assessment in Criminal Sentencing. *Annual Review of Clinical Psychology, 12*, 489–513. https://doi.org/10.1146/annurev-clinpsy-021815-092945

Oleson, J. C. (2011). Risk in Sentencing: Constitutionally Suspect Variables and Evidence-Based Sentencing. *SMU Law Review, 64*(4), 1399–1402.

Rudin, C. Wang, C & Coker, B. (2020). The Age of Secrecy and Unfairness in Recidivism Prediction. *Harvard Data Science Review* 2(1).

Salganik, M. J., et al. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences, 117*(15), 8398–8403. https://doi.org/10.1073/pnas.1915006117

Skeem, J. L., & Lowenkamp, C. (2016). Risk, Race, and Recidivism: Predictive Bias and Disparate Impact. *Criminology, 54*(4), 680–712. https://doi.org/10.1111/1745-9125.12123

Starr, S. B. (2014). Evidence-Based Sentencing and the Scientific Rationalization of Discrimination. *Stanford Law Review, 66*(4), 803–872.

Schwemer, S.F., Tomada, L. & Pasini, T. (2021). Legal AI Systems in the EU's proposed Artificial Intelligence Act. In: *Proceedings of the Second International Workshop on AI and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA 2021)*, held in conjunction with ICAIL 2021, June 21, 2021, Sao Paulo, Brazil.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). *HCR-20: Assessing the risk for violence (version 2)*. Mental Health, Law, and Policy Institute, Simon Fraser University.

Wolff, M. A. (2008). Evidence-Based Judicial Discretion: Promoting Public Safety Through State Sentencing Reform. *New York University Law Review, 83*(5), 1389–1419.

Završnik, A. (2019). Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of Criminology, 18*(5), 1–20. https://doi.org/10.1177/1477370819876762