Hidden Inconsistencies Introduced by Predictive Algorithms in Judicial Decision Making

Travis Greene[1], Ching-Fu Lin[2], Han-Wei Liu[3], and Galit Shmueli[1]
[1] Institute of Service Science, National Tsing Hua University, Taiwan
[2] Institute of Law for Science and Technology, National Tsing Hua University, Taiwan
[3] Department of Business Law and Taxation, Monash University, Australia

Abstract

Algorithms, from simple automation to machine learning, have been introduced into judicial contexts for purposes of increasing consistency and efficiency of legal decision making. In this paper, we describe four types of inconsistencies introduced by risk prediction algorithms. These inconsistencies are the result of the need to operationalize legal concepts and human behavior into specific measures that enable building predictive algorithms. Yet, these inconsistencies are likely to be hidden from their users—judges, parole officers, lawyers, and other decision makers. We describe the inconsistencies, their sources, and propose possible indicators and solutions. We consider the issue of inconsistencies due to the use of algorithms in the larger context of current trends towards more autonomous algorithms and less human-understandable behavioral big data.

I. Introduction

For years, the U.S., and now additional countries, have been applying data analytics and algorithms in the decision-making process of law enforcement agencies, corrections officials, and judges. This trend turns on two interrelated factors: consistency and efficiency (Gertner 2010). First, the data-driven approach appears to represent a *cost-effective* solution, aiding criminal justice officials in prioritizing government resources and in predicting and controlling complex individual behaviors. Second, proponents argue that "evidence-based" tools provided by big data analytics can help reduce human bias and increase consistency, and therefore improve the criminal system (Council of Europe, 2018). Both are key drivers of the move in the U.S. towards a "smarter" regime by incorporating statistical and machine learning (ML) risk prediction algorithms into the decision-making processes of judges (Berk 2015). Machine learning is now increasingly being applied to predict the likelihood of the recidivism or flight of those awaiting trial or offenders in bail and parole procedures (Završnik, 2019). While these new techniques may indeed be useful tools in the judicial context, experts disagree about their appropriateness. Some scholars see the algorithmization of the judicial decision-making roles as inevitable, and argue that courts should embrace autonomation to better serve their mandates (Volokh 2019). Others, however, argue that adopting ML automation tools in the judicial decision-making process may reflect and even exacerbate existing bias and discrimination embedded in the training data (Liu et al. 2019). Still others call for applying well-designed (even if imperfect) algorithms as a way to counter human judges' biases and inconsistencies and hence improve the existing criminal justice system (Corbett-Davies 2017).

Focusing on one of the underlying rationales of applying algorithmic tools in the judicial context—bringing evidence-based, algorithmic "consistency" to the human-centered, error-prone judiciary—this paper argues that an additional layer of inconsistencies may emerge due to how algorithmic tools are designed, employed, and communicated. Current literature has already identified inconsistencies between different statistical fairness criteria (Courtland, 2018). Here we identify and unfold four further types of hidden inconsistency problems that may apply to both risk prediction algorithms as well as identification algorithms. Along these lines, we argue that *consistency*, broadly defined here as treating similar cases similarly, can be compromised and caution is therefore warranted when adopting algorithmic tools. More importantly, as jurisdictions increasingly apply not only hard coded rules and statistical models but also blackbox machine learning algorithms, such as Deep Learning, to the court system, we will need to

reconsider the use of ever-more fine-grained behavioral big data to facilitate algorithmic classifications and predictions in legal settings and its implications for the future of the judiciary.

II. Inconsistencies Introduced by Algorithms

Algorithmic risk prediction models introduce new inconsistencies into judicial decision-making, yet judges, lawyers, and defendants may be unaware of their impact. Several of these inconsistencies arise not from the final application of the algorithms themselves, but rather from the human elements involved earlier in the data gathering, cleaning, and variable selection phases; these human factors include the data subjects, data collectors, data scientists, and other human decision makers. Critical choices are made by the data scientists developing the system who must navigate between abstract legal and behavioral concepts and precise measurable data and mathematical quantities. Finally, further inconsistencies stem from the way in which the algorithms' performance is evaluated and communicated to end users. Table 1 summarizes the human elements involved in each of the four inconsistencies, which we describe next.

Table 1: The human elements involved in each of the four inconsistencies

Inconsistency In...	Human Elements
1. Choice of Measured Outcome/Predictors	Data scientists and Data Engineers
2. Choice and Quality of Training Dataset	Data scientists, Data Subjects, Data Collectors
3. Predictive Accuracies of Subgroups	Data scientists, Data Subjects, Data Collectors
4. Communicated Risk Scores	Data scientists, Judicial Decision-makers

Inconsistency One: Choice of Measured Outcome and Predictors

The fluid nature of legal concepts is often at odds with the data-specific requirements of statistical and machine learning algorithms. The performance of a predictive model heavily depends on the choice of the measurement designated as "outcome to predict." Designing a predictive model entails searching for an algorithm that best predicts that measurement for new individuals. Hence, different choices of an outcome measurement can introduce dramatic inconsistencies in predicted scores. In their examination of bias in predictive algorithms, Barocas & Selbst (2016) warn, "Danger resides in the definition of the class label itself and the subsequent labeling of examples from which rules are inferred." Legal labels such as "violent" or "low-risk" can be operationalized in a variety of ways, each way leading to different predictions.

Judicial decision-making is the process of how judges come to a solution for a given issue through legal reasoning, interpretation, and application based on the relevant laws, regulations, precedents, facts as well as the guiding values and policy orientations of the judges. In the context of sentencing, judicial decision-making is often concerned with assessing an individual's likelihood of future unlawful behavior, given his unique personal characteristics, social connections, and past history. Ideally, sentencing and treatment determinations are made in order to minimize the probability of future unlawful behavior. These decisions, however, are unlike those found in engineering contexts, where outcomes of interest can be objectively determined and procedures for achieving them can be mathematically formalized and optimized. In predictive decision-making scenarios faced by judges, for example, the operationalization of legal concepts such as recidivism is not straightforward and a variety of measures are equally valid.

Surprisingly, there is no generally accepted legal definition of "recidivism." Some define it as the duration between two events,[1] while others measure it by a dichotomous "reconvicted/not" within a certain time period from some event (see Table 2). Brennan et al. (2009) define it as "a finger-printable arrest

[1] e.g. "Days from release date to the point of the first … warrant date" (Breitenbach et al. 2010)

involving a charge and a filing for any uniform crime reporting (UCR) code." A 2012 U.S. Department of Justice Special Report[2] uses four measures of recidivism: rearrest, reconviction, resentence to prison, and return to prison with or without a new sentence within a three-year period following the prisoners' release, further distinguishing between "in-state" and "out-of-state" recidivism. The much-publicized ProPublica study of bias in risk assessment tools defined it as a new arrest within two years of the original crime for which the subject was assessed, discounting any minor offenses and municipal ordinance violations.[3] However, this definition is problematic because it counts subjects who were arrested but were not convicted, and those whose charges were dropped. The choice of recidivism measure also leads to different selected models and algorithms: predicting the *expected time until recidivating* calls for a different type of model than for the *probability of recidivating in the next five years*. The same model cannot produce both.

Table 2: A wide variety of criteria are used to measure "recidivism" by different systems

Criteria	**Options used by different studies / systems**
Events	Arrest, conviction, incarceration
Degree	Felony, misdemeanor, public ordinance
Time periods	2 years, 3 years, 5 years,...
Since	Previous crime, incarceration, arrest, conviction
Inclusion criteria	In-state / out-of-state
Predicted outcome type	time-to-recidivate, probability of recidivism, hazard ratio

Changing the time horizon in the definition of recidivism dramatically changes the base rate of the phenomenon, which in turn affects relevant predictive performance measures. For example, Rice & Harris (1995) showed that changes in the definition of"violent recidivism" to include any new violent crimes within a horizon of 3.5, 6, and 10 years (resulting in base rates of 15%, 31%, and 43%, respectively), caused the VRAG (Violent Risk Appraisal Guide) model's sensitivity, positive predictive power, Chi-squared statistic, and odds ratio to fluctuate dramatically. Nevertheless, end users of these predictive tools may not know which definition of violent recidivism was used in the final model. A judge might incorporate the VRAG's risk assessment score in passing down a sentence, assuming the score reflect the defendant's risk of recidivating within the next two years, when in fact it actually reflects a ten-year risk.

A second source of inconsistency arises due to selecting predictors—the measures used as inputs into the predictive algorithm, which are predictive of the outcome. Developing predictive models requires trading off model complexity with model fit. An overly-complex model that contains many predictors can be made to fit arbitrarily well to a given dataset, but in doing so loses the ability to predict well for new, unseen individuals. A key data science strategy towards achieving this tradeoff is the process of "model selection"—selecting a subset of predictors. Cheng (2009) argues that model selection is analogous to the "reference class problem" in legal risk-assessment, because the predictors included and excluded in the final model determine the criteria that potentially define one's reference group. The more predictors selected, the narrower the potential scope of an individual's reference group used to calculate his risk. Unfortunately, there is no simple solution for deciding an individual's most relevant reference class, which is arguably a legal and moral issue. The model selection process nevertheless creates inconsistencies between and across individuals and systems that rely on different sets of predictors. For example, the most accurate model for predicting Parolee A's recidivism might exclude predictors legally-relevant in predicting Parolee B, who rightfully belongs to a different set of reference classes. Prior choices of which data to

[2] https://www.bjs.gov/content/pub/pdf/rpr94.pdf
[3] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

collect will determine which predictors, and therefore which reference classes, can result after the model selection process. Yet this adds inconsistency into judging risk, both between and across individuals.

Finally, data scientists may add inconsistency when attempting to improve model performance. For example, an outcome measure that includes rare crimes might create an unbalanced dataset that is difficult to model; the data scientists might thus group together the set of rare crimes.[4] Breitenbach et al. (2010) conclude, "The performance of any instrument will vary depending on the population (e.g., prison releases or probation cases), measurement error in the scale or outcome, and type of outcome (e.g., arrests versus violations or returns)." For example, for "violent" felony crimes, a crude dichotomous outcome measure hides important causal distinctions (drug-related vs. domestic violence) and leads to inconsistent risk predictions not appropriately linked to the context of the individual under assessment (Breitenbach 2010).

Potential Indicators & Solutions: Ideally, there should be agreed upon legal definitions of terms such as recidivism and what constitute valid and reliable measures. The field of psychometrics can provide scientific guidelines as to constructing and evaluating suitable measures. Identifying inconsistencies due to various outcome measures requires transparency by system designers, as well as sensitivity analyses when using alternative measures. The data science approach of ensembles (averaging multiple models) with different recidivism measures is another possibility for enhancing consistency, although how to ensemble such different models is not straightforward. For model selection, sensitivity analysis and ensembles may again be useful approaches. Another approach empowers judges as users: some systems (e.g. the Oregon state judicial support systems) allow the *interactive* selection of a relevant reference class, effectively allowing manual model selection for each individual case.

Inconsistency Two: Choice and Quality of Dataset Used to Train the Algorithm

Algorithmic decision-making tools range in their reliance on data for developing the algorithm. At one extreme,"hard-coded" rules, such as "exceeding the speed limit by over 10% leads to a $100 fine," require no data. On the other extreme, "autonomous" machine learning algorithms, such as those used in recommendation systems or deep learning algorithms, "learn" the entire structure between input and output from data they were trained on. Predictive models used in judicial contexts, however, currently occupy a middle ground that relies on statistical models. Models at this level must still rely on many data science choices, such as *which* phenomena are of interest to study and how best to measure or aggregate them. Together, these choices result in different training datasets and therefore different predictive models. Hence, when using a dataset to train an algorithm that will later be used to predict new people's behaviors, *measurement error* and *selection bias* can lead to inconsistencies in resulting scores.

Datasets used for training and evaluating a model's predictive power must include the outcome measure of interest as well as a set of relevant input measures. These inputs and outcomes are linked through the an algorithm to form a predictive relationship. For recidivism, data on offenders' arrests are needed. These data, however, suffer from noisy measurement and can even be intentionally-manipulated either by the data subjects or by the data collectors. Plea-bargains, for example, introduce noise when a person is arrested for one crime but ultimately charged for another, less serious one (Breitenbach 2010). As another example, pre-parole questionnaires are used in risk assessment systems in the Pennsylvania Corrections Department (Berry-Jester 2015). Parolees are asked simple yes/no questions regarding their past behaviors (e.g. whether they have ever had a drug or alcohol problem). But parolees are never asked to indicate the *severity* of the problem or what exactly constitutes a drug or alcohol problem.

[4] The judicial decision support system used in the state of Oregon advises judges to increase the breadth of crimes, age, and even gender in order to collect enough "similar" cases that can be used.

Measurement error can arise also due to manipulation of human data collectors through financial incentives. Organizational pay-for-performance schemes may motivate police to incorrectly classify crimes and describe arrests (Muller 2018). Such gaming techniques used by police to "hit their targets" include: a) "choosing not to believe complainants," b) "recording multiple incidents in the same area as a single crime," and c) "downgrading incidents to less serious crimes." Maltz (1999) also mentions the decades' old FBI *hierarchy rule* for ensuring that crimes would not be double counted. The rule states that when two different types of crime occur in the same incident, only the category of the most serious crime is counted.[5] Such gaming strategies and arbitrary operating procedures lead algorithms to learn relationships more reflective of record-keeping limitations and wishful thinking than of reality.

Data used to train and test predictive algorithms for use in criminal law often suffer from over-representation of some populations and under-representation of other populations. Such misrepresentation leads to inconsistencies when deployed to new members of under-represented populations, or even more extremely, to completely different populations. For example, this problem is particularly acute when predictive models largely trained on male inmates are applied to female inmates (Hannah-Moffat & Shaw 2001). Systems developed and trained for one application are now used not only for new populations (geographically, demographically, temporally,[6] etc.) but also in originally-unintended contexts, such as an algorithm developed for decision on prison releases being applied for probation decisions.[7] For example, the Public Safety Assessment tool was trained on pre-trial data from 300 U.S. jurisdictions, but is applied statewide in Kentucky, Arizona, New Jersey, and Utah where pre-trial populations are likely different in terms of ethnic subgroups from the overall population of jurisdictions.[8]

Data collection mechanisms are one cause of such mis-representation. For example, approximately 10% of randomly-selected inmates declined to participate in the data collection efforts for the COMPAS Reentry risk assessment tool (Breitenbach et al. 2010). Combining these data with arrest data generated by predictive policing algorithms introduces a further selection bias: Increased police surveillance leads to more arrests and recorded criminal incidents. When such data are then fed into judicial decision making algorithms, they create a self-sustaining feedback loop that reflects more the nature of the crime sampling process than actual patterns of criminal behavior.[9] In addition, data science practices of "data preparation" can lead to selection bias, for example, by removing records with missing values. When such missingness is systematic (e.g. refusal to respond to sensitive questions about criminal behavior), it leads to exclusion of specific populations.

Potential Indicators & Solutions: Identifying and tackling measurement error requires transparency and knowledge about data collection practices[10] and mechanisms for testing data quality. To identify selection bias, rather than relying on reported performance of an algorithm, it should be evaluated on the new population and context in which it is planned to be deployed. There are some reports of such testing,

---

[5] "If a convenience store robbery results in the death of the store clerk, this would be classified as a homicide rather than a robbery--because homicide is a more serious crime than robbery" (Maltz 1999).

[6] "Crime is a normative phenomenon, that is, it depends on human values, which change over time and place. Algorithmic calculations can thus never be accurately calibrated given the initial and changing set of facts or 'reality'." (Završnik 2019)

[7] COMPAS was originally designed as a risk assessment software to help corrections agencies manage cases, but is now used for sentencing and other purposes (Kehl et al. 2017).

[8] https://www-cdn.law.stanford.edu/wp-content/uploads/2019/05/PSA-Sheet-CC-Final-5.10-CC-Upload.pdf

[9] According to a recent report by the Partnership on AI, "arrest, conviction, and incarceration data are most appropriately viewed as measures of *official response* to criminal behavior," not *criminal behavior* as such. www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system

[10] The Stanford Pretrial Risk Assessment Tools Factsheet Project audits several pre-trial risk prediction tools, providing details about the training and testing data used, predictive performance in terms of AUC, and other important factors such as which jurisdictions use it and whether its users must undergo prior training.

especially in new geographies.[11] Further, for predictive models trained using over or undersampling of rare or overrepresented classes (e.g., women in COMPAS Reentry), it is especially important to assess predictive performance in a context as close as possible to the actual one.

Inconsistency Three: Predictive Accuracy and Precision for Different Subgroups

No matter the algorithm used, some individual or group-level records are harder to predict than others, resulting in larger prediction errors. Our notion of inconsistency reflects a similar problem. Cases for which there is substantial uncertainty in individual predictions should not be treated the same as those for which there is little uncertainty, yet many current systems fail to indicate this. One factor that can affect such differences in prediction errors is the amount of data collected on some subgroups relative to others. Similarly, another factor concerns predictor information that might have more predictive power for some subgroups compared to others. For example, the nature of the predictor "criminal history" might be more predictive of recidivism for older defendants than for very young ones without a detailed criminal history. As the U.S. Sentencing Commission (USSC) notes, "an individual offender's criminal record cannot decrease with age, only stay constant or increase."[12] Yet algorithms will always produce a predicted score, and the predicted score will not convey this uncertainty. In the machine learning literature, the idea that false positive rates should be relatively balanced across protected classes of persons (e.g., race or religion) is called "equal opportunity" (Hardt et al. 2016).[13] However, detecting unequal predictive accuracy in subgroups requires access to the performance evaluation information and knowledge of predictive metrics. Judges without training in machine learning are unlikely to be aware of such disparities across subgroups.

Potential Indicators & Solutions: Identifying different levels of predicted accuracy for different subgroups can be done at the performance evaluation stage or at a later auditing stage that has access to the testing data. Ideally, variation should be reported along with predicted scores (e.g. "for people with this type of input data, the algorithm performs poorly").[14] Such discoveries can also lead to improving the model to better predict poorly-predicted subgroups. Machine learning approaches such as boosting can help train models improve predictive accuracy for "difficult-to-predict" individuals.

Inconsistency Four: Communicated Risk Scores

Current judicial risk prediction tools use algorithms that produce *class-based,* as opposed to *individual-based* risk probabilities. However, what is often conveyed to the end user—the judge—can be considered a *dichotomous prediction* with *qualified confidence risk levels* (Grisso & Appelbaum 1992). These typically take the form of statements such as, "The defendant will commit a similar act in the future with high/medium/low risk." Risk levels are created by the tool designer by grouping probabilities into a few "risk levels" in order to "aid practitioner interpretation" (Chiappa & Issac 2019). In some U.S. states, such as Kentucky, practitioners are required to do this.[15] Indeed, converting probabilities into risk levels can increase consistency across judges who might interpret probability in different ways, but this is at the cost of precision (validity). Yet a more precise and finer breakdown of risk levels may limit the discretion enjoyed by a judge in the judicial decision-making process. For example, Gastwirth (1992) reports a study of judges in the Eastern District of New York revealing varying interpretations of probability assigned to important legal standards of proof, such as "preponderance" or "beyond a reasonable doubt." Despite legal standards

[11] NorthPointe reported testing COMPAS "across multiple jurisdictions and state agencies" (Brennan, Dieterich & Ehret 2009; Breitenbach, Dieterich, Brennan & Fan 2009 – In press); Tollenaar & van der Heijden (2019) trained algorithms on Dutch conviction data, then tested it on North Carolina prison data.

[12] https://www.ussc.gov/sites/default/files/pdf/research-and-publications/research-publications/2017/20170309_Recidivism-CH.pdf

[13] However, "equal opportunity" defined this way may be "statistically impossible to reconcile if there are differences across two groups – such as the rates at which white and black people are being rearrested" (Courtland 2018).

[14] This is recommended by the Partnership on AI. www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system

[15] www-cdn.law.stanford.edu/wp-content/uploads/2019/05/PSA-Sheet-CC-Final-5.10-CC-Upload.pdf

being phrased in probabilistic sounding terminology, they are only roughly translatable into numerical quantities. Importantly, when an algorithm's predicted probabilities are converted to risk levels, this process is neither driven by legal considerations nor aligned with the specific chosen outcome (that is, the outcome measure is not "risk level"), thereby creating an inconsistency across systems and applications.

We note that probabilities can potentially be converted into dichotomous predictions, which is the case in automated decision-making systems. However, in systems aiding human decision making, providing probabilities allows the decision maker to integrate risk information with additional information. Ashworth & Zedner (2014) explain that predicted probabilities "avoid the problem of false positives because they do not claim to foretell whether a given offender will or will not offend, but only to estimate a given population's propensity for violence." Nevertheless, there may still be a large, unaccounted for gap between a predicted probability and dichotomous prediction. For example, an algorithm might classify an individual as a recidivist if their predicted probability ranges between .50 to .99.

Potential Indicators & Solutions: Algorithm training and performance must be aligned with the deployment context in a legal setting.[16] For systems required to report only risk levels, model development, comparison and performance evaluation should all be geared towards and calibrated for accurate prediction of those risk levels, rather than probabilities. At the same time, risk levels should be guided by legal considerations, and tested in legal settings to assure proper interpretation.[17]

III. Discussion

We have outlined four inconsistencies created by algorithms used for risk prediction in judicial decision-making. For each, we pointed out potential indicators as well as potential solutions. These inconsistencies arise from the inherent difficulty of translating legal and behavioral criteria into a precise mathematical framework (Gastwirth 1992), which is required for deploying predictive algorithms. Current risk models are mostly limited to statistical regression models, which are interpretable and computationally stable. However, it is possible that the next generation of risk models will rely on more autonomous data-driven machine learning algorithms, such as those used in predictive policing and tax fraud detection. If things do indeed move in this direction, the four hidden inconsistencies we mentioned above will persist.

In fact, a move towards more data-driven machine learning methods may lead to additional inconsistencies. Analogous experience with personal data in the EU (the GDPR) and the financial services industry (Basel II/III) suggests boundaries will need to be drawn regarding the appropriate level of granularity of personal data and also the complexity and interpretability of machine learning methods used in the legal context. Classification and regression trees are a likely first move towards machine learning algorithms in the judicial decision-making context, due to their transparency and interpretability. Yet, they may exacerbate the "reference class problem," where predictions for the same person would rely on different granularities of categorization (e.g., some models will predict based on a reference class of "all males older than 20 years," while another may use "all males younger than 30, in neighborhood A, with two school-age children"). More data-driven machine learning methods are also less stable than regression models. Re-running the algorithm with the same data but a different initialization random seed can give a different output. Even using different software can lead to a different result. As different tools rely on different algorithms, legal definitions, and data collection mechanisms, inconsistency in an individual's predicted scores, interpretability, and predictive accuracy is likely to grow.

[16] "At what probability of recidivism should a prisoner be granted parole? Whether this threshold ought to be a 40 percent or an 80 percent risk of recidivism is an inherently 'political' decision based on the social, cultural and economic conditions of the given society." (Završnik 2019)

[17] "There is still a need to ascribe meaning to the probability and 'translate' it in the context of the judicial setting." (Završnik 2019)

A second possible advance in the next generation of risk models is the move from using only curated and well-understood input measures from dedicated datasets that correlate with the outcome behavior of interest, towards using a broad range of "features" derived from fusing multiple data sources (e.g. social media), with data in multiple formats (numerical, text, image, video, network, etc.). Such a shift is already underway in other risk-modeling industries, such as auto insurance, where policy rates are now being determined through behavioral driving data. In China, using such behavioral big data is already common in law enforcement and tightening up the government's social and political control. This fundamentally contrasts with the contemporary understanding of the rule of law (Chen et al. 2018) which is largely reflected in the legal systems of most Western democracies. For one, courts are concerned with the relevance of evidence as well as with ascertaining it was properly obtained (Gastwirth 1992). For another, algorithmic consistency may not be aligned with the court's mandate to ensure nondiscriminatory equal protection, due process, and transparency and accountability.

It is unclear whether these legal standards will continue to apply to data and algorithms used in risk assessment tools. Experience in the financial and insurance industries suggests a more pragmatic "big data" approach may ultimately prevail. Nevertheless, caution is warranted when employing predictive algorithms in the context of courts. The inconsistencies we described will only become more difficult to identify and solve. We therefore recommend that, in addition to existing calls for proper transparency and accountability mechanisms (Kroll et al. 2017), only well-designed, thoroughly and regularly tested algorithmic tools are permitted in high-stakes settings.

IV. References

Ashworth, A., & Zedner, L. (2014). *Preventive justice*. OUP Oxford.

Berk, R., & Hyatt, J. (2015). Machine learning forecasts of risk to inform sentencing decisions. *Federal Sentencing Reporter*, 27(4), 222-228.

Breitenbach, M., Dieterich, W., Brennan, T., & Fan, A. (2010). Creating Risk-Scores in Very Imbalanced Datasets: Predicting Extremely Violent Crime among Criminal Offenders Following Release from Prison. In *Rare Association Rule Mining and Knowledge Discovery: Technologies for Infrequent and Critical Event Detection* (pp. 231-254). IGI Global.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (pp. 3315-3323).

Chen, Y. J., Lin, C. F., & Liu, H. W. (2018). 'Rule of Trust': The Power and Perils of China's Social Credit Megaproject. *Columbia Journal of Asian Law*, 32(1).

Cheng, E. K. (2009). A practical solution to the reference class problem. Colum. L. Rev., 109, 2081.

Gastwirth, J. L. (1992). Statistical reasoning in the legal setting. *The American Statistician*, 46(1), 55-69.

Grisso, T., & Appelbaum, P. S. (1992). Is it unethical to offer predictions of future violence?. *Law and Human Behavior,* 16(6), 621-633.

Hannah-Moffat, K., & Shaw, M. (2001). *Taking risks: Incorporating gender and culture into the classification and assessment of federally sentenced women in Canada*. Ottawa, Ontario: Status of Women Canada.

Kehl, Danielle, Priscilla Guo, and Samuel Kessler. 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.

Kroll, J. A., Huey J., Barocas S., Felten E. W., Reidenberg J. R., Robinson D. G., Yu H. (2017). Accountable algorithms, *University of Pennsylvania Law Review*, 165, 633-705.

Muller, J. Z. (2018). *The tyranny of metrics*. Princeton University Press.

Starr, S. B. (2015). The new profiling: Why punishing based on poverty and identity is unconstitutional and wrong. *Federal Sentencing Reporter*, 27(4), 229-236.

Završnik, A. (2019). Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of Criminology*, 1477370819876762.