12-2022

Forks over knives: Predictive inconsistency in criminal justice algorithmic risk assessment tools

Travis GREENE

Galit SHMUELI

Jan FELL

Ching-Fu LIN

Han-wei LIU
Singapore Management University, hanweiliu@smu.edu.sg

Citation

GREENE, Travis; SHMUELI, Galit; FELL, Jan; LIN, Ching-Fu; and LIU, Han-wei. Forks over knives: Predictive inconsistency in criminal justice algorithmic risk assessment tools. (2022). *Journal of the Royal Statistical Society: Statistics in Society Series A*. 185, (2), S692-S723.
Available at: https://ink.library.smu.edu.sg/sol_research/4397

Received: 24 July 2021 | Accepted: 14 September 2022

DOI: 10.1111/rssa.12966

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Series A Statistics in Society

Forks over knives: Predictive inconsistency in criminal justice algorithmic risk assessment tools

Travis Greene[1] | **Galit Shmueli**[1] | **Jan Fell**[1] | **Ching-Fu Lin**[2] | **Han-Wei Liu**[3]

[1]Institute of Service Science, National Tsing Hua University, Hsinchu, Taiwan

[2]Institute of Law for Science and Technology, National Tsing Hua University, Hsinchu, Taiwan

[3]Department of Business Law and Taxation, Monash University, Clayton, Victoria, Australia

Correspondence:
Travis Greene, Institute of Service Science, National Tsing Hua University, Hsinchu, Taiwan.
travis.greene@iss.nthu.edu.tw

Funding information
Taiwan National Science and Technology Council, Grant/Award Numbers: 108-2410-H-007-091-MY3, 111-2628-H-007-001

Abstract

Big data and algorithmic risk prediction tools promise to improve criminal justice systems by reducing human biases and inconsistencies in decision-making. Yet different, equally justifiable choices when developing, testing and deploying these socio-technical tools can lead to disparate predicted risk scores for the same individual. Synthesising diverse perspectives from machine learning, statistics, sociology, criminology, law, philosophy and economics, we conceptualise this phenomenon as *predictive inconsistency*. We describe sources of predictive inconsistency at different stages of algorithmic risk assessment tool development and deployment and consider how future technological developments may amplify predictive inconsistency. We argue, however, that in a diverse and pluralistic society we should not expect to completely eliminate predictive inconsistency. Instead, to bolster the legal, political and scientific legitimacy of algorithmic risk prediction tools, we propose identifying and documenting relevant and reasonable 'forking paths' to enable quantifiable, reproducible multiverse and specification curve analyses of predictive inconsistency at the individual level.

KEYWORDS

algorithmic risk prediction, criminal justice, forking paths, multiverse analysis, pluralism, predictive inconsistency, specification curve analysis

1 | INTRODUCTION: THE EVOLUTION OF ALGORITHMIC RISK ASSESSMENT INSTRUMENTS

Prediction and classification methods have played a role in criminal justice for over a century. For nearly as long, doubts regarding the suitability and performance of such methods have loomed. In 1895, for instance, Francis Galton wondered whether different judges might impose different patterns of penalties for the same kinds of offenders (Gottfredson, 1987). What started as 'behavioral forecasts' informing parole decisions in the 1920s (Berk & Bleich, 2014) eventually evolved into more sophisticated 'actuarial' prediction methods used since at least the 1970s in the United States. Since then, data-driven statistical and algorithmic risk assessment instruments (ARAIs) have been used in a variety of criminal justice contexts including sentencing, pretrial incarceration, parole, probation supervision levels and security levels decisions (Brennan, 1987). Predictive algorithms are now deployed at numerous stages in the judicial process, from pre-guilt fact-finding to post-guilt sentencing stages, as part of a general trend towards 'evidence-based sentencing' (McKay, 2020). Although the acronym ARAI originally referred to 'actuarial risk assessment instrument', it now also refers to more sophisticated 'algorithmic' tools (Partnership on AI, 2016).

In an era of budget cuts and limited resources, US-based judges have increasingly become 'managerial', relying on the power of information systems to quickly resolve disputes and organise growing caseloads (Resnik, 1982). Consider the CompStat (computerised statistics) system used by police departments around the world since the 1990s (Bratton & Malinowski, 2008). CompStat and other tools like it combine data-driven decision-making with private sector 'performance management' techniques emphasising standardised metrics to track progress and promote accountability. But lunches are not free. The growing pressures of bureaucratic efficiency and consistency—resulting in determinate sentencing laws, and, in some jurisdictions, the mandatory use of risk assessment tools (Stevenson, 2018)—have left judges with varying degrees of discretion in overriding the tool's recommendation (Garrett & Monahan, 2020). Still, the standardised pre-sentence reports compiled by probation officers have been criticised by some as a 'ritual' used to 'maintain the myth of individualized justice' (Rosecrance, 1988). More recently, the 2016 *State v. Loomis* case re-ignited similar debates when Loomis unsuccessfully argued in the Wisconsin Supreme Court that including a prediction from the COMPAS ARAI in his pre-sentencing investigation report violated his rights to due process and individualised sentencing (Liu et al., 2019).

Despite these concerns, the digitisation of court records, new forms of data collection, and cheaper processing power have catalysed the use of data analytics and algorithms in the decision-making processes of law enforcement agencies, corrections officials, and judges (Coglianese & Ben Dor, 2021). Today, over 200 ARAIs are used worldwide, and US corrections agencies alone rely on over 19 different tools (Duwe & Rocque, 2017; Fazel & Wolf, 2018). Proponents of modern data-driven approaches to risk assessment emphasise their cost-effectiveness in prioritising limited government resources and in predicting and controlling individual behaviour. ARAIs promise to reduce human bias and inconsistency, and provide a scientific and evidence-based approach to the judicial process (Završnik, 2020). Data-driven risk assessment helps manage the complexity of judicial decision-making, which often involves estimating the likelihood of future unlawful behaviour, given an individual's unique personal characteristics, social connections, and past history (Monahan & Skeem, 2016). In US pre-trial contexts, risk assessment tools are part of reformatory efforts aimed at reducing wealth-based disparities

and providing 'equal justice for poor and rich, weak and powerful', in line with constitutional guarantees to equal protection and due process (Mayson, 2017). The overall goal is to move the United States toward a 'smarter' regime by incorporating statistical and machine learning (ML) risk prediction algorithms into the criminal justice system (Berk & Hyatt, 2015). Although the United States is unique in its reliance on and early adoption of ARAIs, many countries are now implementing predictive algorithms in their criminal justice systems, including China (Li, 2020) and the United Kingdom (Cui, 2020). Others, such as Austria, Latvia and the Netherlands, are debating their use (Council of Europe, 2019).

Experts fall into several camps about the appropriateness of ARAIs in criminal justice decision-making contexts. Some scholars view the algorithmisation of judicial decision-making as inevitable and argue courts should embrace automation (Volokh, 2018). Well-designed (even if imperfect) algorithms can mitigate human judges' biases and inconsistencies (Završnik, 2020), improve the criminal justice system (Corbett-Davies et al., 2017), and reduce prison and jail populations. ARAIs can help limit unwanted judicial discretion in sentencing by giving offenders with similar characteristics similar sentences, resulting in more consistent and rational sentencing policy (Frase, 2000). As Eckhouse et al. (2019) explain, 'With the (ARAI) scores as guidance, judges and policymakers can apply the same model to every case and claim they have used an objective, neutral mechanism of fair treatment'.

Yet critics allege ARAIs are simply a new mode of 'penological control' (Feeley & Simon, 1992) adapted from managerial methods of risk assessment. Instead of rehabilitating, reintegrating or retraining offenders, these methods merely shuffle the allocation of 'risky' offenders in society by 'selective incapacitation' (Auerhahn, 1999). And many scholars, particularly from science and technology studies, are skeptical of claims that the rationality and objectivity of big data justifies the use of ARAIs in criminal justice contexts (boyd & Crawford, 2012; Dressel & Farid, 2018; Moses & Chan, 2014). Complex and proprietary algorithms may inadvertently reflect and exacerbate existing social biases and discrimination embedded in training data (Liu et al., 2019), thus reproducing injustices already present in society (Mittelstadt et al., 2016). A recent and related stream of research draws from surveillance studies, critical race theory, sociology and feminist theory to argue that ARAIs express a form of 'technoscientific' power whose ahistorical and ostensibly neutral algorithms, data, and computer code serve to maintain structural inequalities across racial and social boundaries (Benjamin, 2019; D'ignazio & Klein, 2020; Eubanks, 2018).

1.1 | The role of prediction in punishment and sentencing

Whether algorithmic risk assessment is considered a boon or bane may depend on one's theory of criminal punishment. Modern ARAI proponents often assume a utilitarian or consequentialist ethics in which public safety and deterrence is the primary goal (see, e.g., Corbett-Davies et al., 2017). Sentencing and punishment thus assume a forward-looking perspective with a focus on minimising the costs and probability of future unlawful behaviour (Hamilton, 2015). Indeed, some US states, such as Pennsylvania, require judges to consider defendants' future dangerousness during sentencing (Berk & Bleich, 2014). Efficient gains in public safety result from replacing 'individualized diagnosis and response' with 'aggregate classification systems' used for 'surveillance, confinement, and control' (Feeley & Simon, 1992). Yet the utilitarian, consequentialist rationale for ARAIs worries legal scholars who see their use as conflicting with individual constitutional rights to avoid self-incrimination and receive procedural due process and equal protection (Starr, 2014, 2015).

In the instrumental pursuit of efficiency, predictive algorithms now carry out the 'selective incapacitation' policies popular in the United States from the 1980s and 1990s (Kehl & Kessler, 2017). Selective incapacitation 'selects' or classifies high risk individuals using statistical methods in order to isolate them from the social community (Blackmore & Welsh, 1983). The method relies on finding highly correlated indices of a behaviour of interest, particularly those which reliably discriminate between low, medium and high risk persons. The ends of public safety are believed to justify the predictive means. Berk and Bleich (2014) note, '(if) shoe size is a useful predictor of recidivism, then it can be included as a predictor … Why shoe size matters is immaterial'. The utilitarian believes the net benefits in reduced crime from identifying and incapacitating high-rate offenders outweigh the costs of 'wasted imprisonment' on low-rate offenders (Moore et al., 1984). Utilitarians are generally willing to accept more false positives (falsely imprisoning an innocent person) and fewer false negatives (treating a dangerous criminal leniently) to achieve greater public safety.

In contrast to utilitarian and consequentialist theories of punishment, retributivists look backwards. They argue that if punishment serves to inflict retribution, confer just desert, or express moral condemnation, basing it on prediction is fundamentally misguided. Indeed, *just desert* theories preclude justifying punishment for its instrumental value, such as improving public safety (Darley et al., 2000). Just punishment is therefore *proportional* to the seriousness of the committed crime, or in the words of philosopher Immanuel Kant, to the 'internal wickedness' of the perpetrator (Carlsmith et al., 2002). Retributivists disapprove of algorithmic prediction because it violates the presumption that one must actually do something prohibited by law—not merely be 'at risk' of doing something—before losing one's right to determine one's own future (Feinberg, 1970). That is, moral blame stems from one's past choices, not one's potential future conduct (von Hirsch, 1984). In general, retributivists claim algorithmic approaches conflict with individual rights and moral autonomy, and criticise the narrow focus on predictive accuracy in legal decision-making (Pundik, 2008; Underwood, 1979; Wasserman, 1991).

1.2 | Contribution and scope

The contribution of this paper is fourfold. First, we collect and organise a diverse literature of ARAI-related issues from ML, statistics, sociology, criminology, law, economics, philosophy and related fields. Second, we unite these sources under our newly introduced umbrella concept of *predictive inconsistency*. Predictive inconsistency occurs when an algorithmic system generates disparate predicted scores for the same individual, based on conceptually-justified-but-technically-different choices by ARAI designers. Third, we identify and taxonomise sources of predictive inconsistency and relate them to real-world data science and criminal justice choices. Fourth, drawing on extant legal practices and the normative framework of scientific and political pluralism, we propose adding multiverse and specification curve analysis techniques to ARAI development and auditing toolkits in order to estimate a lower bound on predictive inconsistency, as well as to illuminate diverse sources of discretionary bias and quantify their impact on predicted risk scores.

The paper proceeds as follows. Section 2 considers the nature and value of consistency in various fields and relates legal consistency to our guiding concept of predictive inconsistency. Section 3 describes the ARAI-building process. Section 4 identifies and discusses specific sources of predictive inconsistency related to decisions at the various stages of ARAI development and deployment. Section 5 considers how more complex predictive algorithms combined with new

forms of behavioural data may exacerbate predictive inconsistency. To quantify and evaluate predictive inconsistency, Section 6 draws on pluralist scientific and political theories to propose a publicly-justified 'multiverse' prediction framework for documenting and reproducing forking paths. Section 7 provides our conclusions.

2 | CONSISTENCY IN ALGORITHMIC RISK ASSESSMENT: FORKS VERSUS KNIVES

Below we describe several notions of consistency and relate them to concepts of justice in order to set the stage for our new term, *predictive inconsistency*. In logic and philosophy, consistency is a property of a set or system of axiomatic propositions, all of which, if true, render the conclusion true (Marcus, 1980). In scientific modelling, a valid model makes predictions consistent with observed data (Oreskes et al., 1994). In statistics, consistency is a property of an estimator, while in psychometrics, the reliability of a measurement is its consistency across occasions or across items designed to measure the same construct (Groves et al., 2011). Moving to law, witness testimony and forensic evidence are said to be consistent with a legal theory explaining the facts of the case (Kiely, 2005). In jurisprudence, consistency denotes coherent judicial decisions (Sunstein et al., 2002), and can refer to consistency of approach or consistency of sentences (Lovegrove, 1997). Consistency of approach signifies that sentences are related to penal aims and case facts in a principled way, while consistency of sentences implies like sentences for like cases. Real-world examples of consistency in US law are captured by principles such as *procedural due process* and *procedural regularity*, which protect against arbitrary and biassed lawmaking targeting specific individuals or groups (Kroll et al., 2017).

2.1 | Legal consistency

While consistency rests in a dense web of relations to other concepts, we focus on its relation to justice. Early Greek philosophers believed the ideal society, as well as the cosmos, to be ruled by an intelligible, rational and eternal 'law and order' best expressed through numerical relations (Kline, 2012). Justice involves acting to restore this order or proper balance. Aristotle, for instance, believed the just person develops a disposition for moral discernment through habitual exercise of practical reasoning to particular cases in a variety of contexts (Ackrill, 1988).

Aristotle's insights are still relevant to judicial reasoning today. Judges must engage in a delicate balancing act that involves inferring judicial principles from individual cases in a consistent way while also maintaining the flexibility to handle new and unexpected circumstances (Cane, 2002). Just legal judgments are proportionally balanced in the sense of not being under- or over-fit to the case at hand; this balance provides the desirable property of generalising to future cases. Yet as Aristotle also insisted, one cannot give a priori rules specifying precisely where this balance must lay (Ackrill, 1988). As in predictive modelling, the price of rigid consistency with past data (i.e., memorisation) is generalisability. Justice requires consistency, but not absolutely: 'law never requires a judge to sacrifice "justice" on the altar of consistency' (Cane, 2002, pg. 20). Still, consistent legal decision-making is valuable, even if not absolutely. Consistency may be evidence of the underlying accuracy and objectivity of legal judgments (Legomsky, 2007), as when multiple judges decide similar cases in similar ways. The consistency of law also provides citizens with a degree of predictability to guide them in planning and orienting their lives (Hart, 1961).

This aspect of consistency is related to the concept of the *rule of law* (Rosenfeld, 2000) which, in part, makes citizens accountable only to those laws which are publicly promulgated and capable of being followed.

Perhaps the most worrying effect of slavish adherence to consistency is its tendency to blind one to novelty and difference, the recognition of which can expand the scope and application of law, as often occurs in landmark cases. For instance, different judges may examine what appears to be superficially the 'same' type of case and come to contradictory conclusions, yet this disagreement does not always imply judicial inconsistency. A highly experienced judge may have noticed a small but relevant detail—perhaps 'hidden' under the surface—that had escaped the attention of other judges. Once identified, this newly discovered relevant difference (Burch & Furman, 2019) can modify pre-existing rules or principles, or be used to craft new ones, thereby contributing to the evolution and expansion of law and concept of justice (Cane, 2002; Hayek, 1973).

In practice, however, it is difficult to separate mere difference from inconsistency. Justice means different things to people of different religions and worldviews, and so legal disputes may themselves concern the question of what constitutes a relevant moral or legal difference (Rawls, 2005). For this reason, Perez (2006) argues that perfect legal consistency—a perfectly uniform application of rules—in a diverse, 'pluralistically sensitive' society would result in systemic injustice to certain groups whose moral and political beliefs may conflict with those of legislators and judges. Perfect legal consistency can, for instance, be achieved by repressing freedom of belief and conscience, or by restricting opportunities for good faith dialogue with others. In contrast, some degree of legal inconsistency should be expected and even desired in a pluralistic society, as its existence and eventual resolution hints at the possibility of developing more comprehensive legal principles, moral concepts and, ultimately, progress in the pursuit of justice.

2.2 | Predictive inconsistency

Although we argue some degree of legal inconsistency in pluralistic societies should be expected and even tolerated, inconsistency in science is generally viewed as undesirable. Inconsistency in decision-making results in 'noise' and derives from variation over occasions, variation across individuals or both (Kahneman et al., 2016). Indeed, the inability to reproduce experimental findings has prompted a methodological 'crisis' in psychology (Simmons et al., 2011). Inconsistent ARAI predictions may invite legal controversy for similar reasons. Even relatively simple 'actuarial' risk assessment tools can generate conflicting risk scores for the same individual. For example, in the 2018 case *State v. Gordon*, Gordon was classified in his pre-sentencing report as *high-risk* for sexual re-offense using the three-category SOTIPS tool, while the five-category STATIC-99R classified him as *average risk* (Supreme Court of Iowa, 2017).

Motivated by these methodological and real-world legal concerns, we focus on a new term, *predictive inconsistency*. Predictive inconsistency captures conceptually-justified-but-technically-different choices by ARAI designers leading to disparate predicted risk scores for the same individual. Predictive inconsistency arises when conceptually-similar-but-technically-different data science and data collection choices produce disparate and possibly contradictory predictions. Our notion of predictive inconsistency differs from recent work in statistics and ML focusing on the implications of pipeline underspecification (D'Amour et al., 2020), the *Rashomon effect* (Breiman, 2001) and various properties of the *Rashomon set* of 'almost equally-accurate predictive models' (Dong & Rudin, 2020; Fisher et al., 2019; Semenova et al., 2019). In line with legal debates around individualised justice, our approach emphasises

the impact of data science decisions on individual predicted scores, not properties of the models themselves. Second, our proposal aims to examine variability across these predicted scores without necessarily requiring actual (ground truth) outcomes or labels (except at the model training stage), whereas most work in this area computes predictive accuracy or goodness of fit metrics that assume access to actual outcomes.

Predictive inconsistency is important to understand as it affects public-facing algorithms in high-stakes contexts. Statisticians Gelman and Loken (2014) illustrate the problem using Jose Luis Borges' metaphor of a 'garden of forking paths' to describe data analysis decisions and variable human factors they refer to as 'researcher degrees of freedom'. These degrees of freedom represent unacknowledged yet non-trivial sources of uncertainty whose impact is not yet well studied (Sauerbrei et al., 2020). Crucially, these data-dependent choices are not laid out in advance, and researchers often do not realise their particular findings are one outcome of many possible forking paths, each representing a different sequence of data analysis decisions (Dwork et al., 2015; Gelman & Loken, 2014). Data scientists developing ARAIs face similar problems. Because no general algorithm or inductive bias is guaranteed a priori to be the best for a particular problem (Wolpert, 1996), data scientists try a variety of strategies, methods and algorithms—often with multiple tuning parameters—to find a predictive model delivering the best generalisation performance on new, unseen data presumed to come from the same distribution (Duwe & Kim, 2017).

We claim the forking paths problem will likely be compounded as the algorithms and sources of data used in ARAI development increase in complexity, and as the variety of actors and institutions involved grows. Just as scientific publications can gloss over analytical missteps and false starts in retrospectively idealised final analyses (Latour & Woolgar, 2013; Pickering, 1995), a deployed ARAI embodies an idealised *single path* through what is actually a 'garden' of many path-dependent and contingent data science practices, such as unacknowledged and undocumented choices by the human factors involved earlier in generating, gathering and cleaning the data, and later in developing, testing and deploying the algorithm. Yet judges, corrections officers and other end-users of the resulting tool's predictions may underestimate the cumulative impact of these discretionary choices and practices on an individual's predicted risk score.

We believe the socio-technical nature of ARAIs and resource limits of real-world criminal justice systems require us to reconsider the value and nature of predictive inconsistency. We should not expect ARAIs to exhibit the level of consistency provided by the controlled laboratory conditions of the natural sciences or the mechanistic precision of the assembly line. A more realistic expectation—aligned with the actual practices of legal reasoning in a diverse society—is that a well-functioning ARAI should reliably assign an individual into a risk category across a variety of 'reasonable' forking paths, which we interpret as representing *good enough* or *reasonable* rather than *optimal* or *rational* descriptions of the same underlying phenomenon viewed from different vantage points. We think this more inclusive approach can bolster the scientific, legal and political legitimacy of these new and often controversial algorithmic tools.

Our re-framing of predictive inconsistency embraces the democratic spirit of scientific and political pluralism (see e.g., Bohman, 2006; Dewey & Rogers, 2012; Habermas, 2015; Kellert et al., 2006; Kitcher, 2003; Rawls, 2005; Shadish, 1993). Broadly speaking, pluralism endorses tolerating some inconsistency and conflict in the transformative search for overlapping consensus on scientific and political issues by deliberating on diverse yet reasonable worldviews, problem representations, classification schemes, analytical methods, economic and social interests and interpretive assumptions. Pluralism challenges the default assumption that ARAIs can or should embody a single, unique logic of decisions leading to a 'best' model resulting in a

single, clear-cut final prediction ('knives'). Pluralism, in our view, instead engages a diverse team of individuals—in terms of both functional (i.e., cognitive) and identity diversity (Hong & Page, 2004)—in ARAI design, development and auditing focused on assessing the distribution of predictions generated for an individual from a multiplicity of coequally reasonable models ('forks'). Our commitment to democratic pluralism might also warrant the possibility of non-use of an ARAI, particularly when its predictive inconsistency exceeds bounds set by a diverse group of data scientists, domain and legal experts.

3 | BUILDING AN ARAI

ARAIs are socio-technical systems composed of technical artefacts, human agents and resource-constrained institutions whose rules and norms influence and guide the behaviour of human agents (Selbst et al., 2019; van de Poel, 2020). We use the general term *ARAI designers* to refer to institutional stakeholders and industry and/or academic software development and data science teams, often with criminal justice domain expertise. To concretise the socio-technical nature of ARAI design and relate it to data science decisions, we provide a brief description of the ARAI-building process (see Figure 1).

First, a data source must be found, representing a sample from a theoretical target population, such as *all prisoners in Minnesota*. These data are often drawn from court or administrative records (Ritter, 2013). Duwe (2014), for instance, obtained the data for the MnSTARR tool from the Minnesota Correctional Operations Management System (COMS), based on a sample of 11,375 male offenders released from prison between 2003 and 2006. Next, ARAI designers must define and operationalise the outcome or construct of interest, such as recidivism—defined broadly as the commission of subsequent crime within a set time period—and decide on a particular time horizon. Duwe (2014) defines recidivism as *reconviction for a criminal offense within 4 years of release from prison*. Designers also need to decide how many risk categories to include (e.g., low, medium, high) and how much resources they can realistically devote to the different categories.

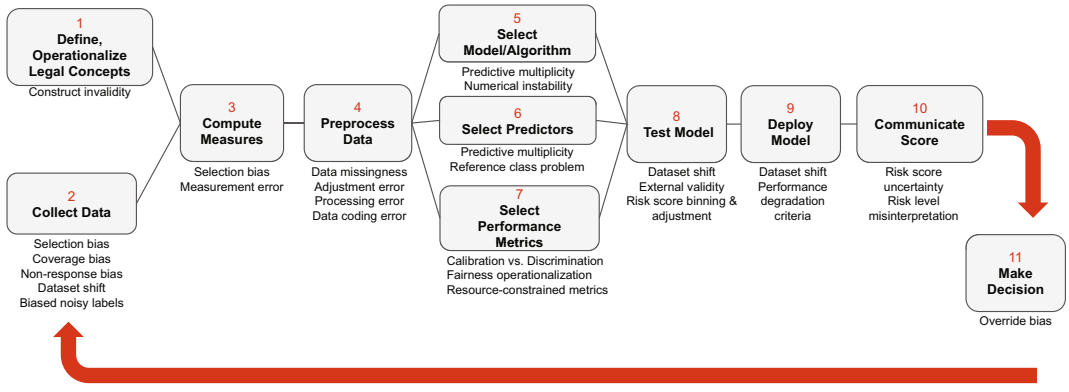

FIGURE 1 Steps and potential sources of predictive inconsistency in the development and use of an algorithmic risk assessment instrument (ARAI). Discretionary 'forking path' decisions by ARAI designers can reduce the tool's ability to provide consistent predictions for the same individual. Common descriptors for the sources of inconsistency are listed below the step. [Colour figure can be viewed at wileyonlinelibrary.com]

Duwe (2014) uses four risk levels with cutoffs varying by gender and type of recidivism (e.g., sexual or violent).

Designers then select predictor variables given the data available and theoretical or policy concerns. Duwe (2014) considered over 100 different predictors, keeping just eight using a bootstrap variable selection procedure. Examples include *prior supervision failures*, *total felony convictions*, *drug offense convictions*, and *false information given to police convictions*. ARAI designers typically also consider the relative costs of misclassification (false positives and negatives).

After building several models and evaluating their resulting predictive accuracy and risk distributions, designers may adjust aspects of the model or data. Assuming accuracy and risk distributions are appropriate given the resource limitations of the institution developing the tool, the model is then deployed for offenders in the agency's caseload. For MnSTARR, male inmates are scored at prison intake and once again prior to their release, after updating 'dynamic factors' that may have changed during incarceration, such as undergoing chemical dependency treatment (Duwe, 2014). This final risk prediction is used to decide the post-release level of community supervision. Lastly, designers monitor the performance over time to ensure its accuracy and reliability.

4 | POTENTIAL SOURCES OF PREDICTIVE INCONSISTENCY

This section describes the real-world process of data collection and algorithmic development and illustrates its impact on predictive inconsistency, as shown in Figure 1. To guide our presentation, we follow a taxonomy of errors affecting data-driven processes commonly used by statisticians and social scientists (Groves et al., 2011). While textbooks tend to focus on issues of *sampling error* and statistical inference, we emphasise the effects of a variety of *non-sampling errors* such as *construct invalidity*, *measurement error*, *processing* and *data coding error*, *coverage bias*, *non-response bias*, and *adjustment error* on predictive inconsistency. Note that discretionary choices in earlier steps can propagate and exacerbate predictive inconsistencies in later steps. An example is measurement error at the data collection stage that later impacts predictors and response variables, thereby affecting a model's predictive performance (Kuhn & Johnson, 2013).

4.1 | Defining and operationalising constructs into empirical measures

The law 'teems with devices that defeat uniformity and predictability' (Easterbrook, 1992). Due to its vast scope of application, law possesses an 'open texture' focused on defining general classes of persons or acts (Hart, 1961). That is, for practical reasons, law cannot in advance define every conceivable criminal act in society and therefore considers only general classes of acts (e.g., murder, battery, fraud) under which particular acts may fall. Legal systems, like most natural systems, are open, not closed, and so legal assertions require an element of intuitive judgment and interpretation that both precludes absolute proof or demonstration and resists explicit automation (Schauer, 2009). In contrast, statistical algorithms must presume the validity of a particular interpretation or application of a general label (i.e., categorical outcome variable) to make inferences and predictions at all (Barocas et al., 2019; Selbst & Barocas, 2018). But whether the resulting

predictions rightfully apply to anything in the 'real world' depends on the degree of correspondence between internal model elements and external objects of interest (Selbst et al., 2019). Crucially, the validity of such correspondence can never be absolutely verified (Oreskes et al., 1994).

The often implicit act of interpretation in ARAI development requires explicitly operationalising theoretical and unobservable legal, social and moral concepts known as *constructs* (Groves et al., 2011). *Operationalising* a construct involves specifying a corresponding set of uniquely specifiable physical or mathematical operations (Bridgman, 1927). Yet assessing the validity of measurements of essentially unobservable constructs is not straightforward (Borsboom et al., 2004). For that reason, a modern and pragmatic view holds that *construct validity* is a holistic, evidential judgment of how well a measure captures its intended construct and supports the interpretation of a score, including action taken on the basis of this interpretation (Messick, 1995). Generally, the higher the stakes of assessment, the more important construct validity becomes (Downing, 2003). One implication is that ARAI designers cannot evade the ethical issues of operationalising 'essentially contested' constructs and values such as *fairness* (Friedler et al., 2021; Jacobs & Wallach, 2021; Mittelstadt, 2019).

The operationalisation of recidivism risk is one such contested and complex process that impacts predictive inconsistency. No generally accepted legal definition of recidivism exists, and the literature devoted to discussing, defining and approximating various operationalisations of recidivism has a long history (Rector, 1958). We note, however, that existing definitions of recidivism share three features. Each definition has a starting event from which the measurement of recidivism commences, for example, release from prison. Second, each definition has a measure of failure following the starting event, for example, a subsequent arrest. Third, each definition has a window of recidivism, that is, a follow-up period within which the offender's behaviour is observed (Zgoba & Dayal, 2015). Table 1 illustrates the diversity of definitions.

According to the US National Institute of Justice, recidivism refers to 'a person's relapse into criminal behaviour, often after the person receives sanctions or undergoes intervention for a previous crime'. Some operationalise it as the duration between two events, for example, days from release date to the point of the first warrant date (Breitenbach et al., 2010). Others measure it by a dichotomous 'reconvicted/not' or 'new arrest/not' within a certain time period from some event (Jones & Sims, 1997; Maxfield, 2005). A Bureau of Justice Statistics report uses four measures of recidivism: *rearrest*, *reconviction*, *resentence to prison* and *return to prison with or without a new sentence within a 3-year period following the prisoners' release*, and further distinguishes between 'in-state' and 'out-of-state' recidivism (Langan & Levin, 2002). The much-publicised ProPublica

TABLE 1 Various criteria used to measure recidivism

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

Note: Definitions obtained from Breitenbach et al. (2010), Blumstein and Larson (1971), Maxfield (2005), Jones and Sims (1997), Langan and Levin (2002), Angwin et al. (2016).

study of bias in risk assessment tools defined it as a *new arrest within 2 years of the original crime for which the subject was assessed, while discounting any minor offenses and municipal ordinance violations* (Angwin et al., 2016). 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.

Different choices can lead to different predictive performance and therefore to inconsistent predicted risk scores. Changing the time horizon in the definition of recidivism changes the *base rate* of the phenomenon (i.e. the underlying proportion in the population of interest), which in turn affects relevant predictive performance measures. For example, Rice and Harris (1995) show 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), causes various performance measures of the Violent Risk Appraisal Guide (VRAG) model to fluctuate dramatically. Nevertheless, end users of an ARAI may not be aware that different definitions might lead to different risk scores for the same individual.

4.2 | Collecting data and computing measures

Statistics and ML textbooks often assume no errors in the outcome labels and thus narrowly focus on issues of sampling error (Hand, 2006). While increasing the sample size can reduce sampling error, many sources of inconsistency mentioned in the ARAI literature are actually instances of *non-sampling error* (Lohr, 2011). Common non-sampling errors that can arise in the data collection step are *selection bias*, *measurement error*, *non-response bias* and *coverage bias*. Data collection and pre-processing procedures are major sources of non-sampling error and are hard to identify and thus control. Merely having 'big data' does not reduce non-sampling error.

The bias and noise added due to non-sampling error can propagate through to later stages of algorithm development (Suresh & Guttag, 2019). Examples of non-sampling error are confusing survey questions, unobserved social and economic pressures influencing persons to respond in systematically different ways, and optical character recognition (OCR) devices generating systematic errors in the transcription of handwritten text.

Coverage bias arises when building predictive models of well-defined target populations based on court and administrative records. Coverage bias relates to mismatches between the *sampled population* and *target population* (Groves et al., 2011), or in ML terms, the algorithmic *development population* and *use population* (Suresh & Guttag, 2019). Depending on what is available, some ARAIs will rely on data from convicted populations, while others on arrested populations (Duwe, 2014; Ritter, 2013). Yet, in the United Kingdom, only about 2% of all crimes ever face sentencing decisions (Ashworth, 2005). And some are arrested without having committed a crime.

Further, poverty and racial disparities in arrests may be reflected in the administrative data used to train risk prediction algorithms, a phenomenon known as 'biased noisy labels' (Fogliato et al., 2020). Rich, well-connected criminals with access to expensive legal representation may be arrested but convicted less often. Computer crimes, identity theft, tax fraud and a variety of other 'low-priority' white collar crimes are notoriously difficult to detect and so may be under-represented if only arrests or convictions are counted (Cole & Smith, 2007). Ideally, data

should be representative of attributes (e.g., 'propensity to recidivate') of the population or construct of interest. To the extent power and wealth disparities, or policing strategies, are reflected in the data, measures based on such data will lead to issues of *construct invalidity* and *measurement error* and actions taken on the basis of the predicted scores should be critically scrutinized.

Selection bias is a concern when relying on non-randomly selected samples to estimate patterns in the population (Heckman, 1979). In ML, the related term *data set shift* describes the situation where the joint distribution of inputs and output differs between the training and test data sets (Quiñonero-Candela et al., 2009). Data sets used to train and test predictive algorithms for use in criminal law can suffer from over-representation of some populations and under-representation of others (Završnik, 2021). Such misrepresentation leads to inconsistencies when deployed to new members of under-represented populations, or even to completely different populations. An example is when predictive models trained on mostly male inmates are applied to female inmates (Hannah-Moffat & Shaw, 2001). Modern ARAIs, however, increasingly create separate predictive models for men and women (Duwe, 2014). Yet systems developed and trained for one target population are sometimes used in other geographical, demographic, temporal and decision-making contexts, such as an algorithm developed for decision on prison releases being applied for probation decisions (Kehl & Kessler, 2017). Berk (2019) lists many other causes of data set shift including *parole board personnel turnover*, *judges losing elections*, *changes in the number of police and prisoners* and *changes in drug markets*, among other causal factors. As an example of potential generalisation issues, the Public Safety Assessment tool used pre-trial training data from 300 US jurisdictions, but was applied statewide in Kentucky, Arizona, New Jersey and Utah, where pre-trial populations of ethnic subgroups are likely different from the overall population of jurisdictions (Stanford Law School Policy Lab, 2019b). An example of good methodological practice comes from Tollenaar and Van Der Heijden (2019), who trained algorithms on Dutch conviction data and evaluated their external validity on a different target population consisting of North Carolina prisoners.

Measurement error is the difference between a measurement and the true value of a quantity (Crowder et al., 2020). It is typically described using *random*, *systematic* or *differential error* models (Luijken et al., 2019). In statistical inference, measurement error can lead to model non-identifiability, blurring the meaningful interpretation of model parameters (Grace, 2016). But measurement error can also amplify predictive inconsistency (Frénay & Verleysen, 2013), as real-world data gathering in criminal justice settings involves noisy measurement, affecting data labels used during ARAI training and later at the time of scoring (Duwe & Rocque, 2017). In plea bargains, for instance, a person is arrested for one crime but charged for a less serious one (Breitenbach et al., 2010). Despite presumed innocence, poorer persons without high-quality legal representation may be more likely to accept plea bargains (Ashworth & Blake, 1996). Due to constraints on legal resources, the practice of prosecutorial discretion resembles 'satisficing' (Albonetti, 1986) and can result in reduced police booking charges (e.g., reducing a felony to a misdemeanor); these reductions may also vary by race and familiarity with the informal workings of the criminal justice system (Albonetti, 1990, 1992).

Financial incentives can contribute to measurement error by influencing the behaviour of human data collectors. Organisational pay-for-performance schemes may motivate police to incorrectly classify crimes and describe arrests. Gaming techniques used by police to 'hit their targets' include 'choosing not to believe complainants', 'recording multiple incidents in the same area as a single crime', and 'downgrading incidents to less serious crimes' (Muller, 2019, p. 128). Maltz (2019) mentions the FBI hierarchy rule to avoid double counting. If two types of crime occur in the same incident, only the category of the most serious crime is counted, for example,

'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, 2019). Such gaming strategies lead algorithms to learn predictive relationships more reflective of record-keeping limitations and wishful thinking than of reality.

Feedback loops in the criminal justice system (Ensign et al., 2018) not only make data collection and causal interpretation of data complex, but they can also affect measurement error. Due to variation in the amount and quality of data collected, some individual or group-level records are harder to predict than others (Rudin et al., 2022), resulting in greater predictive inconsistency. One explanation for this phenomenon touches on larger structural justice issues, as some minoritised subgroups come into contact with administrative agencies more often than others, leading to class imbalances in resulting data sets (Buolamwini & Gebru, 2018; D'ignazio & Klein, 2020). If risk predictions are combined with arrest data generated by predictive policing algorithms, the resulting increased police surveillance can lead to more arrests and recorded criminal incidents although the underlying crime rate has not changed (e.g., Na & Gottfredson, 2013). When such data are then fed back 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 behaviour (Partnership on AI, 2016). Judges, parole and probation officers may also treat those predicted 'high risk' in systematically different ways (Koepke & Robinson, 2018), leading to *differential measurement error*. For example, algorithm predictions in medical settings can lead doctors to undertake different forms of measurement (e.g., self-report of body weight versus physician-directed measurement on a scale) (Luijken et al., 2019).

Differential measurement effects can be related to an individual's race or gender (Mullainathan & Obermeyer, 2017; Suresh & Guttag, 2019), raising social justice concerns and resulting in predictors of unequal predictive power for some subgroups compared to others. For instance, the predictor *criminal history* might be more predictive of recidivism for older defendants than for younger ones without a detailed criminal history, or for those in living or working areas with a history of intense police surveillance (Browne, 2015). Worried by the explosive growth of mass incarceration in America, sociologists have analysed US administrative, survey, and census data between 1969 and 1999 and concluded that incarceration rates for young, low-skilled black men are so high as to resemble a rite of passage into adulthood, on par with college graduation or military service (Pettit & Western, 2004). Criminology research also suggests that incarceration may actually slightly increase one's probability to recidivate (Cullen et al., 2011; Durlauf & Nagin, 2011). Science and technology studies scholars point to these and other examples as evidence for the role of ARAIs in contributing to a *matrix of domination* (Collins, 2002) that reproduces pre-existing social, historical and economic inequities (Benjamin, 2019; D'ignazio & Klein, 2020).

Pre-parole questionnaires used as input to ARAIs are also susceptible to measurement error. They are subject to all the standard sources of measurement error in surveys, including how they are worded, what time and in what setting they are taken, who administers them, and various faulty recall effects (Whittle et al., 2018). As an example, parolees in the Pennsylvania Corrections Department are asked simple yes/no questions regarding their past behaviours, such as whether they have ever had a drug or alcohol problem (Barry-Jester et al., 2015). But they are not asked to indicate the severity of the problem or what constitutes a 'drug or alcohol problem'. Even more, self-reports can be intentionally manipulated by respondents themselves or by data collectors. COMPAS thus embeds a 'data validity check' in self-report questionnaires to identify respondents suspected of 'lying, sabotage, or incoherent responses' (Brennan & Dieterich, 2018). Given the strong incentive offenders have in avoiding a 'high risk' classification, ARAI designers should

periodically assess algorithms and proxy measures for their susceptibility to strategic manipulation, particularly differential gaming strategies by sub-populations. If detected, designers may add randomness or update the algorithm more frequently (Bambauer & Zarsky, 2018). Yet if proxy measures are chosen wisely, gaming behaviour can be made to incentivise the socially beneficial self-improvement of offenders (Kleinberg & Raghavan, 2020).

4.3 | Data pre-processing

Data scientists may increase predictive inconsistency by employing various pre-processing strategies to improve model performance, especially when trying to predict rare events. In social science and psychometrics, these kinds of issues broadly fall under *adjustment error*, *processing error* and *data coding*, including handling missing values (Groves et al., 2011). Although we do not discuss transforming or standardising predictors, these common operations can also introduce predictive inconsistency (Sauerbrei et al., 2020).

Data missingness involves missing values or incomplete data. When such missingness is systematic (e.g. refusal to respond to sensitive questions about criminal behaviour), it may lead to the exclusion of specific populations. Data scientists must then make decisions about whether and how to impute missing values, keep only complete observations, or do nothing. Each decision may have different and unpredictable effects on the final model. 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). Censored data (unobserved outcomes) is a frequent problem in criminal justice settings because those whose outcomes are observed may be systematically different from those whose are not (Berk, 2019). Missing data can cause further inconsistency at prediction time, if the to-be-predicted record has missing predictor values. ML solutions, such as training multiple models with different subsets of predictors (Saar-Tsechansky & Provost, 2007), are another source of inconsistency.

A frequent pattern in ARAI data is imbalanced outcome variables. For example, in a pre-trial risk tool examined by Eckhouse et al. (2019), only 3.8% were arrested for a violent crime and 4.9% failed to appear in court. When the outcome variable is imbalanced, a common pre-processing step is to group together types of rare categories to achieve better balance among the classes. But such grouping comes with a cost. By replacing examples of drug-related or domestic violence with a crude dichotomous measure, such as 'violent' felony crimes, important criminogenic distinctions are obscured and risk predictions may not properly reflect the context of assessment, leading to inconsistent predictive performance (Breitenbach et al., 2010).

If the outcome variable is imbalanced, predictive algorithms can fail to learn to distinguish between the rare and majority classes. Popular preprocessing strategies to deal with this problem include oversampling or undersampling of rare or overrepresented classes (e.g., women in COMPAS Reentry), using cost-sensitive learning, or synthetically generating new records of the rare events (Weiss, 2013). *Discrimination aware* pre-processing can also involve re-labelling or re-sampling/re-weighting methods (d'Alessandro et al., 2017). A simpler method limits the modeled population to a relevant, less-imbalanced subpopulation that contains the class of interest (Weiss, 2013), such as those committing violent crimes. A meta-analysis of 68 studies of violence ARAIs found this strategy improved predictive accuracy (Singh et al., 2011). This approach also finds legal support in Mayson (2017), who argues pre-trial detention may only be appropriate for persons predicted high risk of violent recidivism, not merely any type of recidivism. But doing

this requires precise criteria for deciding which sub-domains count as 'sufficiently interesting'. Different data scientists may come to different conclusions, thus adding a source of inconsistency in the resulting predictions.

4.4 | Variable selection and the reference class problem

Predictive multiplicity occurs when competing models produce conflicting predictions (Marx et al., 2020). One source of such multiplicity (i.e., predictive inconsistency) arises when selecting variables assumed to be predictive of the outcome. Generalisable inductive inference requires developing robust predictive models that trade off model complexity with model fit (the so-called *bias-variance tradeoff*) (Hastie et al., 2009). An overly complex model with many predictors can be made to fit arbitrarily well to a given data set, but in doing so loses the ability to predict well for new, unseen data. To combat this, a key data science strategy is the process of variable selection (Hastie et al., 2009), where predictors based on specified criteria are included or excluded in the model. Examples include lasso and stepwise regression, pruned classification trees and the bootstrap selection procedure used to build the MnSTARR ARAI. Yet small changes in variable inclusion criteria affect which predictors go into the model, thus acting as a source of predictive inconsistency (Heinze et al., 2018).

Variation in variable selection procedures impacts predictive inconsistency in individualised sentencing because different 'evidentiary estimates' could be given to the same person, each on seemingly justifiable grounds (Rhee, 2007). Predictors included and excluded in the final model determine the criteria potentially defining one's reference group and thus one's predicted risk. The more predictors selected, the narrower the potential scope of an individual's reference group used to calculate risk. But the more specific the reference class, the fewer the individuals belonging to the class, thus increasing uncertainty of the risk estimates. To illustrate, we can calculate risk based on gender alone, or on a larger set of predictors such as gender, age and zip code. In the former, the reference class is other people of the same gender, whereas in the latter the reference class is others of the same gender, age, and in the same zip code.

Cheng (2009) argues that variable selection mirrors the *reference class problem*, which occurs when we 'want to assign a probability to a single proposition, X, which may be classified in various ways, yet its probability can change depending on how it is classified' (Hájek, 2007). The 1995 case of *United States v. Shonubi* provides a vivid example of its implication for legal risk assessment, though in a slightly different context from standard ARAIs. Tasked with estimating the unknown amount of heroin Mr. Shonubi—a Nigerian national and US resident—smuggled into JFK Airport on seven previously undetected trips, Cheng (2009) says 'The court could have considered the amount carried by all drug smugglers at JFK, all Nigerian smugglers regardless of airport, or smugglers in general'. Each reference class assignment results in different estimates. The accuracy of these estimates was important because they decided which sentencing guidelines would be applied to his case, potentially adding several extra months or years in prison. The ensuing legal battles incited a lively discussion of when statistical evidence is admissible as 'specific evidence' to be used in an individual's sentencing proceedings (Tillers, 2005).

Prior data collection choices determine the granularity of possible reference classes and the number of possible subsets of predictors evaluated during variable selection, but ideally variable selection should not depend on the particular data at hand (Heinze et al., 2018). Currently, however, different jurisdictions have access to administrative data of varying granularity and quality. For instance, is *nationality* recorded when drug incidents are reported? If not, then it cannot be

used as a predictor. Deciding an individual's reference class is not only a philosophical issue highlighting the indeterminacy (Leiter, 2007) of a priori defining legally relevant differences justifying differential treatment, but also impacts predictive inconsistency.

4.5 | Selecting performance metrics

4.5.1 | Discrimination versus calibration

AUC (Area Under the [ROC] Curve) is a popular performance measure used in ARAIs, yet misunderstanding what it measures can contribute to predictive inconsistency. As Hand (2009) notes, 'choosing a measure which does not reflect [one's aims] could lead to incorrect conclusions'. AUC estimates the probability that a random positive observation (e.g. recidivator) ranks higher than a random negative (non-recidivator) (Flach, 2019). Algorithms with higher AUC scores have better predictive validity: Duwe and Rocque (2017) define an 'adequate' level for ARAIs as 0.70 or higher. But AUC is appropriate only when one's predictive goal is an accurate ranked ordering of risk scores, otherwise known as *discrimination* (Singh, 2013). AUC says nothing about the accuracy of the predicted score. *Calibration*, in contrast, focuses on the estimation of an exact probability (and compares it to the percentage observed in the test set) (Tollenaar & Van Der Heijden, 2019). As Fawcett (2006) explains, an algorithm with high AUC 'need not produce accurate, calibrated probability estimates; it need only produce relatively accurate scores that serve to discriminate [between] positive and negative instances'.The same model may perform better at one or the other task, depending on the nature of the training data and structure of measurement error (Luijken et al., 2019; Whittle et al., 2018). Lastly, AUC comparisons can be misleading when the underlying ROC curves cross, which commonly occurs when comparing multiple classifiers (Hand, 2009).

4.5.2 | Formalising fairness criteria

Problems of fairness in ML have been formulated as constrained optimisation problems (Corbett-Davies et al., 2017; Zemel et al., 2013), which require trading-off predictive accuracy and various formal notions of fairness (Berk et al., 2021). Fairness metrics can focus on either individual or group outcomes and may be used at pre-processing, in-processing or post-processing stages (Berk, 2019). Our discussion mostly centres on post-processing metrics and does not consider those based on 'causal' comparisons of actual with counterfactual outcome distributions (Kusner et al., 2017). We also leave open philosophical questions of whether it is possible or desirable to formalise fairness or justice (Derrida, 1992; Green & Hu, 2018).

Algorithms can have disparities in predictive accuracy for various racial or gender sub-groups, raising questions of bias and unfair discrimination in the data collection process itself (Angwin et al., 2016). For instance, one measure of discrimination compares true positive rates across groups, a fairness metric known as *equal opportunity* (Hardt et al., 2016). But merely observing different rates of predictive accuracy among subgroups is not necessarily evidence of 'unfair' algorithms. Against initial reports of racial discrimination in the COMPAS ARAI (Angwin et al., 2016), the makers of the tool, Northpointe Inc., argued that when base rates of recidivism differ significantly among racial groups, symmetry in the error rates across groups cannot be achieved (Dieterich et al., 2016). This debate highlights whether unequal base rates are better understood

as symptoms or causes of unjust social conditions, conditions in which some social groups are systematically and unjustifiably more likely to be subjected to ARAIs than others. Likewise, such issues cast doubt on the appropriateness of narrowly treating fairness as a property of algorithms, independent of larger social contexts (Selbst et al., 2019).

At the same time, fairness metrics may help identify deeper structural problems at the data collection and operationalisation stages. Eubanks (2018) illustrates this situation using a Pennsylvania-based tool used to predict cases of child abuse. Children in poor families tend to come into contact with public services such as child protective services, Medicaid, and drug and alcohol treatment programmes more often than those from wealthy families, resulting in more extensive data collection for poor families, in terms of both the number of observations and the number of features. In such situations, the ARAI is apt to confuse 'parenting while poor with poor parenting' (Eubanks, 2018, p. 127). The counter-intuitive implication is that ARAIs trained using sparse data from wealthier families may exhibit greater predictive inconsistency, prompting ARAI critics to call for discontinuing their use or delegitimising their predictions. Meanwhile, ARAI proponents could point to the low predictive inconsistency for poor families as a reason to continue using such tools. Yet it seems unfair to subject some families—namely poorer ones—to more algorithmic decision-making and state surveillance and intervention because more and better quality data has been collected on them. In other words, the mere fact of having more data should not count as a *relevant difference* justifying more surveillance.

As one might expect, basing formal predictive measures on contested legal and ethical concepts is difficult, if not impossible. Imposing fairness criteria can itself impede the social goal of minimising expected violent crime because satisfying the criteria can only be achieved by releasing more high-risk defendants (Corbett-Davies et al., 2017). There is thus an inherent tension between treating individuals equally and achieving 'algorithmic fairness' by setting race-specific decision thresholds. Berk (2019) also points out that by not including 'discriminatory' information such as race or gender in the algorithm (i.e., 'fairness through unawareness')—or even their proxies, such as neighborhood—the reduced predictive power may result in greater numbers of dangerous persons released back into communities. Kleinberg et al. (2016) and Chouldechova (2017) demonstrate that in normal situations, no single algorithm can achieve the desired fairness properties of *calibration*, *balance for the negative class* and *balance for the positive class* simultaneously. Only in extremely rare cases of 'perfect risk assignment' by the algorithm and equal base rates among subgroups can a single algorithm satisfy the above-mentioned definitions of fairness. As Dieterich et al. (2016) remark, this is unlikely to hold in practice.

The discussion of competing fairness metrics also highlights a deeper issue: the distinction between individual and group justice (Binns, 2020; Mitchell et al., 2021). Realistically, one must choose the least unfair of a set of unfair algorithms (Speicher et al., 2018). For example, Dwork et al. (2012) base their operationalisation of fairness on an Aristotelian, individualised notion of 'nearest neighbor parity', capturing whether persons represented as similar in predictor space receive similar predictions. Yet many of the approaches detailed above rely on the resulting group-level parity of calibration or false positive rates. Satisfying both forms of fairness requires tradeoffs: minimising between-group unfairness can increase within-group unfairness (Speicher et al., 2018). Consequently, 'hard choices' among fairness metrics cannot be made independently of the complex ethical considerations of individual versus group justice (Binns, 2018b). Equally justifiable metrics can lead to disparate predicted scores. ARAI designers may thus consider generating predicted scores for a variety of fairness definitions and operationalisations based on plurality of legal and philosophical assumptions.

4.5.3 | Lift: resource constrained ranking

Although ARAIs are often justified by claims of cost-saving and efficiency in resource allocation (Duwe, 2014; Ritter, 2013), surprisingly few predictive applications use lift to evaluate classifier performance (Shmueli, 2019). Lift is the ratio of the true positive rate in the *top-n* sample (where *n* is subject to budget or resource constraints) to the true positive rate in the entire test set. The greater the lift, the better the classifier performs compared to random targeting.

In criminal justice contexts, public agencies often have budget and resource constraints limiting their ability to act on every prediction made, a fact with implications for performance evaluation and predictive inconsistency. An ARAI chosen for its superior performance as judged by *capacity-unconstrained* metrics, such as AUC, may be judged inferior when lift is evaluated. As noted above, lift measures a classifier's ranking (i.e. discriminative) ability, just like AUC. But unlike AUC, lift captures the classifier's ability to maximally 'skim the cream' from a subset of individuals (Shmueli et al., 2017, p. 136). The amount of resources that can be invested into acting on the predicted scores decides the size of the subset. As an illustration, Duwe and Kim (2017) note that random forest and logistic regression performance appear very similar in terms of AUC in predicting recidivism, yet random forest much more accurately ranks the highest risk offenders.

In our view, it makes little sense to rely on capacity-unconstrained metrics, such as AUC, in constrained scenarios in which real criminal justice systems operate. Because lift explicitly considers budgetary limits it is well-suited for judicial and policing contexts. But lift requires greater coordination between data scientists and legal domain experts, as they would need information about budgetary constraints and operating scenarios. Lift also raises questions of justice and fairness: is it reasonable to rank people and act only on the top-ranked? When predicted probabilities are—in an absolute sense—low, acting on a predefined top *x%* of cases could violate intuitive notions of fairness.

4.6 | Adjusting and communicating risk scores

ARAI designers convert risk probabilities into risk levels for decision-makers to interpret and act on. Although judges typically have discretion in choosing not to follow ARAI recommendations, evidence from case studies suggests they often prefer to have them if available (Hartmann & Wenzelburger, 2021). An ARAI might display output for an individual similar to the following: 'Risk of Recidivism: Medium; Failure to Appear: Low; Risk of Violence: High' and provide a recommended supervision level (Angwin et al., 2016). ARAIs generally, however, give no indication of the uncertainty associated with predicted risk levels (Partnership on AI, 2016).

Despite little evidence that categorical risk levels promote better informed decision-making (Scurich, 2018), it is common practice to bin continuous probabilities into discrete risk levels in order to 'aid practitioner interpretation' (Chiappa & Isaac, 2018). In some US states, such as Kentucky, practitioners are required to create risk groups (Stanford Law School Policy Lab, 2019a). Nevertheless, there may still be a large, unaccounted for gap between a predicted probability and categorical score. Depending on relative misclassification costs, an algorithm developer might decide to classify an individual as 'high risk' if the predicted probability ranges between 0.50 and 0.99. Mayson (2017) suggests that binning and threshold decisions involving preventative restraint should ideally be based on comparisons with the general population of non-defendants. We note that binning decisions are related to construct validity and measurement error (Hanson et al., 2017).

There are pros and cons to creating risk groups (Hilton et al., 2015). Discretising risk levels discards potentially useful information, but can increase consistency across judges who might interpret raw probabilities in different ways. Gastwirth (1992) cites 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 of evidence' and '[evidence] beyond a reasonable doubt'. On the other hand, a more precise and finer breakdown of risk, such as displaying percentile ranks, absolute recidivism rates, or risk ratios (Hanson et al., 2017), may reduce the consistency of judicial decision-making processes. But access to these more fine-grained distinctions could theoretically promote more individualised sentencing. In either case, when predicted probabilities are converted into risk levels, this choice is usually neither driven by legal or theoretical considerations nor aligned with the outcome measure used in training, which is typically an arrest or conviction, not a 'risk level'. In fact, agencies often adjust algorithms' predictions ex post until predicted risk scores are broadly aligned with the desired or expected distribution of risk in the target population (Ritter, 2013), which is only vaguely estimable. An ARAI designed to help allocate probation supervision levels where 90% of individuals receive a 'high risk' classification is useless in practice. But these ad hoc adjustments and risk groupings create inconsistency across systems and applications.

Finally, as hinted at earlier, the decision to train an algorithm for discrimination or calibration can result in different predictive models being declared most predictively accurate. This is because small differences in relative ranks (e.g., 1–2) may hide large absolute differences in probability (e.g., 0.99–0.01). In other words, the best model for a calibration task may not be the best model for a discrimination task: it will depend on what one is trying to predict. The absolute risks of some crimes may be so low (e.g., a 4.9% failure to appear rate) that converting predicted probabilities into relative ranks or 'high, medium, and low risk-levels' is misleading to end-users (Eckhouse et al., 2019).

5 | POSSIBLE FUTURES FOR ALGORITHMIC RISK ASSESSMENT

Despite methodological debate about the 'exceptional variation' of individual predicted scores and whether this imprecision warrants their disuse (Cooke & Michie, 2010; Imrey & Dawid, 2015), current ARAIs are mostly limited to relatively interpretable and computationally stable statistical regression models. The next generation, however, will likely rely on more automated and flexible ML algorithms with built-in data ingestion and variable selection capabilities (Duwe & Rocque, 2017; Slobogin, 2017), such as those used in predictive policing. We also expect these models to use more granular, micro-level behavioural data on individuals and key criminal justice actors and decision-makers, such as police and judges. As noted earlier, judges have discretion in choosing to follow predicted risk scores, resulting in a 'gap' between algorithmic recommendation and the final judicial decision. Green (2020) details research showing how some judges in Cook County, Illinois are systematically biassed in favour of increased detention, and how defendants' race is correlated with decisions to increase individuals' predicted crime risk. We refer to this phenomenon as *override bias* (see Figure 1, step 11). Further, various institutional and economic pressures—not to mention other psychological factors such as anchoring bias (Chang et al., 2016) and ARAI output misinterpretation (Yacoby et al., 2022)—may affect how and when judges diverge from the algorithm's recommendation. To better understand the relevant factors involved and promote greater accountability, Koepke and Robinson (2018) suggest requiring

judges to justify these divergences and that agencies begin collecting data on when and how they occur.

If a source of data improves the ability to reliably discriminate among risk levels, for-profit companies and under-funded public agencies have clear incentives to use them to more efficiently allocate resources (Berk, 2019). A shift is already underway in some risk-modelling industries such as credit scoring and car insurance, where auto insurance policy rates can be augmented with behavioural driving data and GPS locations. A recent study by Ayuso et al. (2019) used telematic data to design auto insurance policies, emphasising the need to find 'new variables of risk exposure and driver behaviour'. Berk (2019, p. 172) refers to these new individually weak, but collectively powerful sources of predictive information as 'dark structure', arguing 'there is substantial structure in criminal behavior that current thinking in criminal justice circles does not consider'. But if things move in this direction, the hidden sources of predictive inconsistency we mentioned will persist and perhaps become even more difficult to detect.

In terms of ML algorithms, classification and regression trees are a likely next step due to their relative transparency and interpretability. Berk and Bleich (2014) suggest random forests as a candidate ML technique, which also offer predictor importance scores and the ability to easily employ asymmetric misclassification costs in model training (Berk, 2019). Kleinberg et al. (2018) use gradient boosted trees to predict failure to appear cases in New York City. Random forests are used by both the Harm Assessment Risk Tool in the United Kingdom (Oswald, 2018) and a COMPAS tool used by some US prisons for new inmate classification (Brennan et al., 2009). We note, however, that ML algorithms, even tree-based algorithms, can be numerically unstable: re-running the algorithm with the same data but a different random initialisation seed can give a different output (Berk, 2019). Even using different software can lead to a different result, thus increasing predictive inconsistency.

Paralleling the evolution in predictive algorithms, the next generation of ARAIs will rely on new and more finely-grained sources of behavioural data as input (Werth, 2019). As digital era governance and IT infrastructures expand, we foresee a move from using only well-understood input measures (e.g., criminogenic factors), toward 'features' derived from fusing multiple data sources (e.g., social media), with data in multiple formats (e.g., numerical, text, image, video, network). In China, using such behavioural big data is already common in law enforcement and for tightening up the government's social and political control (Jiang & Fu, 2018). In the United States, California has trialed GPS monitoring of high-risk parolees, and in some areas driving under the influence probationers are given remote monitoring anklets (Ridgeway, 2013). In Taiwan, You et al. (2018) designed a portable breathalyser device to reduce drunk driving by logging and analysing complex behavioural data and reporting them to a local probation office. These sources of behavioural data can easily be adapted and included as predictors in future ARAIs. But as the sources and variety of input data grow, so too do worries of robustness under conditions of data set shift (Berk, 2019).

New and more complex sources of data may also conflict with contemporary understandings of the rule of law in Western democracies (Liu et al., 2019). The distinction between using algorithms to explain or predict (Shmueli, 2010) raises legal questions about the relevance and defensibility, accountability, and transparency of high-stakes blackbox predictions when employed by private companies and public authorities (Oswald, 2018). For instance, algorithm audits in the health care domain have uncovered evidence of racial bias due to spurious correlations stemming from construct invalidity (Obermeyer et al., 2019). Unless citizens know they are subjected to such algorithms and they have some way of challenging the results (which itself requires some degree of explainability), predictive algorithms pose the risk of reproducing, not resolving, societal inequities.

6 | TOWARDS ILLUMINATING AND QUANTIFYING FORKING PATHS

As ARAIs use more sophisticated algorithms and harvest more complex sources of behavioural data, how can we illuminate and estimate predictive inconsistency in ways compatible with and justified by legal, political and scientific practices? We suggest a pragmatic compromise based on a philosophy of pluralism. Pluralist ARAI design tolerates some level of predictive inconsistency by viewing a well-functioning ARAI as reliably assigning an individual into a risk category across a variety of 'reasonable' forking paths. This approach aims to cancel out systematic biases of individuals or groups of data scientists by combining predictions made across a variety of agreed-upon reasonable forking paths constituting the design, decision, representational and attentional biases of data scientists from diverse walks of life. Inspired by the 'notice and comment' rulemaking procedures of US government agencies (Mulligan & Bamberger, 2019), the democratic pluralism we endorse insists that deliberation on the 'reasonableness' (Binns, 2018a) of a path be responsive to public criticism from those likely to be subjected to predictions (or their chosen appointees). In this way, individuals can influence decisions affecting their interests (Fung, 2013), or at least be heard in a due process. The predictions generated through the democratic and technologically aided procedure we propose may better achieve the normative legitimacy characteristic of 'procedurally fair' legal judgments (Habermas, 2015; Tyler, 2006). We thus advocate for ARAI design based on publicly contestable, multi-path 'forks' rather than monological, single-path 'knives'.

But not all forking paths are under the control of ARAI designers. For example, data scientists have little to no control over the quality of the administrative data, or how corrections officers conduct interviews and assessments. Further, judges are not required to follow the ARAI's recommendation and may even misinterpret its output. Local legal requirements also constrain construct operationalisation. Yet many paths are under the designers' control. We therefore distinguish between forking paths that are quantifiable by ARAI designers and those that are not. Quantifiable paths comprise a variety of generic data science decisions, which can vary according to local legal and data set-driven considerations. These paths can be used to help illuminate the various sources and impact of predictive inconsistency on individual risk prediction scores.

The first step is to identify and document the multiplicity of paths involved in the development of an ARAI, as outlined in Section 4. We then envision turning identified forking paths into self-contained, shareable, searchable and reproducible computational environments using tools and workflows adapted from software engineering (Forde et al., 2018; Ragan-Kelley and Willing, 2018). Model-level 'predictive multiplicity' metrics (Marx et al., 2020) can be combined with auditing tools such as *data sheets* (Gebru et al., 2021) and *model cards* (Mitchell et al., 2019) to help document data and modelling choices relevant to a specific prediction. The goal of this step is to facilitate internal and third-party ARAI auditing and possibly even foster civic engagement by creating publicly accessible ARAI registries, as several major cities have done (Johnson, 2020).

After identifying and documenting candidate paths, our pluralist approach calls for a diverse group of domain experts to deliberate on which paths are 'reasonable', drawing on their combined technical, legal and theoretical knowledge. Again, any agreed upon definitions, ranges or thresholds should be responsive to and contestable from those likely to be subjected to predictions, or their chosen representatives (i.e., a public data science advocacy group). Table 2 provides a basic template for evaluating the reasonableness of a currently deployed ARAI or one in development.

TABLE 2 Sample template for deliberating on the reasonableness of quantifiable 'forking path' choices by ARAI designers

Step in ARAI development	Data science choices (examples)	Evaluating reasonableness of path
Defining and operationalising constructs	Event type, time period, geographic area	Check local law and domain knowledge (e.g., criminology)
Collecting data and computing measures	Data scientists generally have less control over data collection and quality	Use or create data sheet. Check for: coverage bias, measurement error and selection bias
Data pre-processing	Data imputation, over/under sampling, grouping rare categories, binning continuous features, removing outliers, transforming predictors	Use model cards and exploratory data analysis. Investigate reasons for missingness and/or class imbalance
Model/variable selection	Regression model, random forest, stepwise procedures, bootstrap, regularisation, add pairwise interactions	Examine interpretability and numerical stability (e.g., different random seeds and train/test splits)
Selecting performance metrics	AUC, Brier scores, lift, fairness metrics	Determine: calibration or discrimination goal? Institutional budget constraints? Individual or group fairness?
Adjusting risk scores and communication	Create discrete risk bins	Check local law for binning guidance and budgetary limits

Assuming consensus has been reached on a set of reasonable forking paths, and these paths have been identified and documented in reproducible environments, the next step is generating predictions for individuals. We propose adapting the methods of *multiverse analysis* (Steegen et al., 2016) and *specification curve analysis* (Simonsohn et al., 2020), a technique giving a high-level, visual overview of multiverse analyses. Multiverse analysis stems from applications in psychology (Simmons et al., 2011) and the analysis of experimental fMRI data, where up to 35,000 different forking paths may be at play (Carp, 2012).

Although multiverse analyses generally focus on parameter estimation and statistical inference, we instead suggest adapting these techniques to individual-level predictions for a specific ARAI under development or audit. The essential idea is to evaluate subject-level predictive inconsistency by generating reasonable forking paths and obtaining the set of resulting prediction scores for each subject. The score distribution for an individual is then plotted using specification curves. During ARAI development and auditing, individual-level score distributions can be used for estimating overall predictive inconsistency stemming from manipulable forking paths and for researching sources that might be eliminated. Further, the score distributions can also be used at the time of decision-making to, for instance, illuminate person-specific predictive inconsistency

levels and help end-users determine the appropriate level of credibility to attach to the ARAI's score. Similar to the suggestion of Marx et al. (2020), when an ARAI exhibits an unreasonable degree of predictive inconsistency, we might choose not to make predictions for that person, or implement the model at all.

To reiterate, because of our focus on individual-level prediction scores, the key modifications to multiverse analysis include (1) creating a predictive inconsistency holdout set, ideally based on a purposive or representative sample of the population of interest located where the ARAI is likely to be deployed; (2) determining 'reasonable' minimal predictive and fairness performance metric thresholds, based on deliberation among data scientists, domain and legal experts (e.g., whether AUC, Brier scores or lift should be used and their acceptable thresholds or ranges; which fairness metric is used, etc.); (3) adapting specification curves to display subject-level predicted scores. Specification curves helpfully reveal both model specification details (i.e., which forking paths went into an individual's prediction) and corresponding predicted scores. We envision the results of these analyses open to public comment and scrutiny.

Multiverse analysis can also help to align ARAI design and auditing with principles and practices of legal responsibility aimed at preventing and repairing harms in society. In civil law, for instance, domain specific, community-based standards of 'reasonable conduct' and care are used to assess responsibility for harmful outcomes (Cane, 2002). When evaluating negligence claims in domains such as medicine, a common test for establishing causation entails asking whether, 'but for the negligence of the defendant, the plaintiff would not have been injured' (Epstein, 1973). A further consideration is whether abnormal or deviant conditions were present (Hart & Honoré, 1985). An ex post multiverse analysis can thus help reveal that an individual's predicted risk scores would likely have fallen into some range, 'but for' a particular data science decision. With the help of a diverse community of domain experts to specify 'reasonable' forking paths, multiverse analysis can encourage socially accountable ARAI design by providing standards to assess the reasonableness of a particular design choice. Besides providing useful 'contrastive' or counterfactual causal explanations (Miller, 2019) of predicted risk scores, multiverse analysis can guide efforts to improve data collection and quality and indicate construct invalidity when irrelevant factors have an unexpected and disproportionate impact on an individual's predicted risk score.

Prediction-oriented multiverse analysis is limited by applying only to quantifiable forking paths amenable to manipulation by data scientists. Analyses may therefore only provide a lower bound for an ARAI's predictive inconsistency for a given individual. Privacy concerns, strategic gaming considerations and issues of intellectual property law may also restrict the feasibility and generality of this collaborative approach. Citizens and ARAI end-users might also lack the requisite data science and legal knowledge or motivation to participate in developing community-based standards of reasonableness. For this reason, a specialised government agency could be tasked with setting such standards (see, e.g., Scherer, 2015). In any case, we encourage interdisciplinary research on not only the sources of predictive inconsistency affecting ARAIs and similar socio-technical systems, but also on the larger legal and institutional decision-making processes in which they are embedded.

7 | CONCLUSION

Although controversial, ARAIs increasingly influence decision-making in criminal justice systems worldwide. Predictive inconsistency captures the conceptually-justified-but-technically-

different decisions by ARAI designers that reduce the ability to provide consistent predictions for the same individual. Key sources of predictive inconsistency include construct invalidity, measurement error, processing and data coding error, coverage bias, non-response bias, and adjustment error. These errors relate to the way ARAI designers define, operationalise, collect and preprocess data, select predictor variables, choose models and performance metrics, test and deploy them, and communicate risk scores in many and often conflicting ways. Grounded in the normative framework of scientific and political pluralism, we propose multiverse and specification curve analyses as steps towards reproducible, explainable and democratically accountable ARAI development and auditing. These methods reveal how reasonable deviations (forks) from a single development path (knives) affect the variability of a particular prediction. They also provide new means for contesting predictions and advancing ARAIs' scientific, political and legal legitimacy.

Criminal justice systems operate with limited resources, and ARAIs promise more efficient and consistent decisions. Yet these practical benefits must be weighed against the potential harms of unjust treatment stemming from algorithmic blindness to morally relevant differences in individual persons and cases. Questions of whether and how ARAIs can be used in a just, responsible, transparent and innovative manner will not be resolved by more sophisticated technology, more rigorous mathematical formalism, or more abstract ethical principles alone. Tolerating some degree of predictive inconsistency may be the price we must pay to live in a democratic, dynamic and pluralistic society where legal and social practices, moral concepts, technologies, policing and carceral strategies and even the definitions of crimes evolve over time and are subject to the changing needs and interests of citizens. Whether the resulting imprecision negates the ostensible social and legal utility of ARAIs we alone cannot say. This is largely a political issue (Wong, 2020) that requires not only an appeal to technical expertise, but to normative authority as well. Ultimately, however, we support adjusting the tools and techniques of predictive modeling to fit the complex and plastic nature of persons, law and society, rather than adjusting persons, law and society to fit the idealised and rigid nature of predictive modeling.

ACKNOWLEDGEMENTS

We thank Mark Shope for his valuable comments and feedback and the Associate Editor and three reviewers for their many thought-provoking suggestions. Shmueli and Greene were partially funded by Taiwan National Science and Technology Council (Grant 108-2410-H-007-091-MY3). Lin was partially funded by Taiwan National Science and Technology Council (Grant 111-2628-H-007-001).

DATA AVAILABILITY STATEMENT

Data sharing is not applicable as no new data generated.

ORCID

Travis Greene https://orcid.org/0000-0003-4487-0529

REFERENCES

Ackrill, J.L. (1988) *A new Aristotle reader*. Princeton: Princeton University Press.

Albonetti, C.A. (1986) Criminality, prosecutorial screening, and uncertainty: toward a theory of discretionary decision making in felony case processings. *Criminology*, 24, 623–644.

Albonetti, C.A. (1990) Race and the probability of pleading guilty. *Journal of Quantitative Criminology*, 6, 315–334.

Albonetti, C.A. (1992) Charge reduction: an analysis of prosecutorial discretion in burglary and robbery cases. *Journal of Quantitative Criminology*, 8, 317–333.

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016) *Machine bias*. Available at: www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [Online, Accessed 24th February 2022].

Ashworth, A. (2005) *Sentencing and criminal justice*. New York: Cambridge University Press.

Ashworth, A. & Blake, M. (1996) The presumption of innocence in English criminal law. *Criminal Law Review*, 306–317.

Auerhahn, K. (1999) Selective incapacitation and the problem of prediction. *Criminology*, 37, 703–734.

Ayuso, M., Guillen, M. & Nielsen, J.P. (2019) Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data. *Transportation*, 46, 735–752.

Bambauer, J. & Zarsky, T. (2018) The algorithm game. *Notre Dame Law Review*, 94, 1.

Barocas, S., Hardt, M. & Narayanan, A. (2019) *Fairness and machine learning*. Fairmlbook Available from: http://www.fairmlbook.org

Barry-Jester, A, Casselman, B. & Goldstein, D. (2015) *The new science of sentencing: should prison sentences be based on crimes that haven't been committed yet?* Available from: https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing [Online, Accessed 12th May 2021].

Benjamin, R. (2019) *Race after technology: abolitionist tools for the new JIM code*. Cambridge: Polity Press.

Berk, R. (2019) *Machine learning risk assessments in criminal justice settings*. New York: Springer.

Berk, R. & Bleich, J. (2014) Forecasts of violence to inform sentencing decisions. *Journal of Quantitative Criminology*, 30, 79–96.

Berk, R., Heidari, H., Jabbari, S., Kearns, M. & Roth, A. (2021) Fairness in criminal justice risk assessments: the state of the art. *Sociological Methods & Research*, 50, 3–44.

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

Binns, R. (2018a) Algorithmic accountability and public reason. *Philosophy & Technology*, 31, 543–556.

Binns, R. (2018b) Fairness in machine learning: lessons from political philosophy. In: *Conference on fairness, accountability and transparency*. New York City, NY: PMLR, pp. 149–159.

Binns, R. (2020) On the apparent conflict between individual and group fairness. In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. Barcelona, pp. 514–524.

Blackmore, J. & Welsh, J. (1983) Selective incapacitation: sentencing according to risk. *Crime & Delinquency*, 29, 504–528.

Blumstein, A. & Larson, R.C. (1971) Problems in modeling and measuring recidivism. *Journal of Research in Crime and Delinquency*, 8, 124–132.

Bohman, J. (2006) Deliberative democracy and the epistemic benefits of diversity. *Episteme*, 3, 175–191.

Borsboom, D., Mellenbergh, G.J. & Van Heerden, J. (2004) The concept of validity. *Psychological Review*, 111, 1061.

boyd, D. & Crawford, K. (2012) Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15, 662–679.

Bratton, W.J. & Malinowski, S.W. (2008) Police performance management in practice: taking compstat to the next level. *Policing: A Journal of Policy and Practice*, 2, 259–265.

Breiman, L. (2001) Statistical modeling: the two cultures. *Statistical Science*, 16, 199–231.

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*. Hershey, PA: IGI Global, pp. 231–254.

Brennan, T. (1987) Classification: an overview of selected methodological issues. *Crime and Justice*, 9, 201–248.

Brennan, T. & Dieterich, W. (2018) Correctional offender management profiles for alternative sanctions (compas). In: Singh, J.P., Kroner, D.G., Wormith, J.S., Desmarais, S.L. & Hamilton, Z. (Eds.) *Handbook of recidivism risk/needs assessment tools*. New York: John Wiley & Sons.

Brennan, T., Dieterich, W. & Ehret, B. (2009) Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior*, 36, 21–40.

Bridgman, P.W. (1927) *The logic of modern physics*. New York: Macmillan.

Browne, S. (2015) *Dark matters*. Durham, NC: Duke University Press.

Buolamwini, J. & Gebru, T. (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In: *Proceedings of the Conference on fairness, accountability and transparency*. New York City, NY: PMLR, pp. 77–91.

Burch, M. & Furman, K. (2019) Objectivity in science and law: a shared rescue strategy. *International Journal of Law and Psychiatry*, 64, 60–70.

Cane, P. (2002) *Responsibility in law and morality*. Oxford: Hart.

Carlsmith, K.M., Darley, J.M. & Robinson, P.H. (2002) Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, 83, 284.

Carp, J. (2012) On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments. *Frontiers in Neuroscience*, 6, 149.

Chang, Y.C., Chen, K.-P. & Lin, C.-C. (2016) *Anchoring effect in real litigation: an empirical study*. University of Chicago Coase-Sandor Institute for Law & Economics Research Paper.

Cheng, E.K. (2009) A practical solution to the reference class problem. *Columbia Law Review*, 109, 2081.

Chiappa, S. & Isaac, W.S. (2018) A causal Bayesian networks viewpoint on fairness. In: *IFIP international summer school on privacy and identity management*. New York: Springer, pp. 3–20.

Chouldechova, A. (2017) Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data*, 5, 153–163.

Coglianese, C. & Ben Dor, L. (2021) Ai in adjudication and administration. *Brooklyn Law Review, Forthcoming, University of Pennsylvania School, Public Law Research Paper*.

Cole, G. & Smith, C. (2007) *The American system of criminal justice*. Belmont, CA: Thomson. Wadsworth Publishing.

Collins, P.H. (2002) *Black feminist thought: knowledge, consciousness, and the politics of empowerment*. New York: Routledge.

Cooke, D.J. & Michie, C. (2010) Limitations of diagnostic precision and predictive utility in the individual case: a challenge for forensic practice. *Law and Human Behavior*, 34, 259–274.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. & Huq, A. (2017) Algorithmic decision making and the cost of fairness. In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. Halifax, NS: pp. 797–806.

Council of Europe. (2019) *Practical examples of ai implemented in other countries*. www.coe.int/en/web/cepej/practical-examples-of-ai-implemented-in-other-countries [Online, Accessed 20th May 2021].

Crowder, S., Delker, C., Forrest, E. & Martin, N. (2020) *Introduction to statistics in metrology*. New York: Springer.

Cui, Y. (2020) *Artificial intelligence and judicial modernization*. New York: Springer.

Cullen, F.T., Jonson, C.L. & Nagin, D.S. (2011) Prisons do not reduce recidivism: the high cost of ignoring science. *The Prison Journal*, 91, 48S–65S.

d'Alessandro, B., O'Neil, C. & LaGatta, T. (2017) Conscientious classification: a data scientist's guide to discrimination-aware classification. *Big Data*, 5, 120–134.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A. et al. (2020) Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*.

Darley, J.M., Carlsmith, K.M. & Robinson, P.H. (2000) Incapacitation and just deserts as motives for punishment. *Law and Human Behavior*, 24, 659–683.

Derrida, J. (1992) Force of law: the mystical foundation of authority. In: Cornell, D., Rosenfield, M. & Carlson, D. (Eds.) *Deconstruction and the possibility of justice*. New York: Routledge, pp. 3–67.

Dewey, J. & Rogers, M.L. (2012) *The public and its problems: an essay in political inquiry*. University Park, PA: Penn State Press.

Dieterich, W., Mendoza, C. & Brennan, T. (2016) *Compas risk scales: demonstrating accuracy equity and predictive parity*, Vol. 7. Northpointe Inc.

D'ignazio, C. & Klein, L.F. (2020) *Data feminism*. Cambridge, MA: MIT Press.

Dong, J. & Rudin, C. (2020) Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence*, 2, 810–824.

Downing, S.M. (2003) Validity: on the meaningful interpretation of assessment data. *Medical Education*, 37, 830–837.

Dressel, J. & Farid, H. (2018) The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4, eaao5580.

Durlauf, S.N. & Nagin, D.S. (2011) Imprisonment and crime: can both be reduced? *Criminology & Public Policy*, 10, 13–54.
Duwe, G. (2014) The development, validity, and reliability of the minnesota screening tool assessing recidivism risk (mnstarr). *Criminal Justice Policy Review*, 25, 579–613.
Duwe, G. & Kim, K. (2017) Out with the old and in with the new? an empirical comparison of supervised learning algorithms to predict recidivism. *Criminal Justice Policy Review*, 28, 570–600.
Duwe, G. & Rocque, M. (2017) Effects of automating recidivism risk assessment on reliability, predictive validity, and return on investment (ROI). *Criminology & Public Policy*, 16, 235–269.
Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O. & Roth, A. (2015) The reusable holdout: preserving validity in adaptive data analysis. *Science*, 349, 636–638.
Dwork, C., Hardt, M., Pitassi, T., Reingold, O. & Zemel, R. (2012) Fairness through awareness. In: *Proceedings of the 3rd innovations in theoretical computer science conference*. Cambridge, MA, pp. 214–226.
Easterbrook, F.H. (1992) Abstraction and authority. *The University of Chicago Law Review*, 59, 349–380.
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, 185–209.
Ensign, D., Friedler, S.A., Neville, S., Scheidegger, C. & Venkatasubramanian, S. (2018) Runaway feedback loops in predictive policing. In: *Proceedings of the conference on fairness, accountability and transparency*. New York City, NY: PMLR, pp. 160–171.
Epstein, R.A. (1973) A theory of strict liability. *The Journal of Legal Studies*, 2, 151–204.
Eubanks, V. (2018) *Automating inequality: how high-tech tools profile, police, and punish the poor*. New York: Martin's Press.
Fawcett, T. (2006) An introduction to roc analysis. *Pattern Recognition Letters*, 27, 861–874.
Fazel, S. & Wolf, A. (2018) Selecting a risk assessment tool to use in practice: a 10-point guide. *Evidence-Based Mental Health*, 21, 41–43.
Feeley, M.M. & Simon, J. (1992) The new penology: notes on the emerging strategy of corrections and its implications. *Criminology*, 30, 449–474.
Feinberg, J. (1970) *Doing & deserving; essays in the theory of responsibility*. Princeton: Princeton University Press.
Fisher, A., Rudin, C. & Dominici, F. (2019) All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20, 1–81.
Flach, P. (2019) Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward. In: *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33, pp. 9808–9814.
Fogliato, R., Chouldechova, A. & G'Sell, M. (2020) Fairness evaluation in presence of biased noisy labels. In: *International conference on artificial intelligence and statistics*, pp. 2325–2336.
Forde, J., Head, T., Holdgraf, C., Panda, Y., Nalvarete, G., Ragan-Kelley, B. et al. (2018) Reproducible research environments with repo2docker. In: *ICML reproducibility in machine learning workshop*. Stockholm.
Frase, R.S. (2000) Is guided discretion sufficient-overview of state sentencing guidelines. *The Saint Louis University Law*, 44, 425.
Frénay, B. & Verleysen, M. (2013) Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25, 845–869.
Friedler, S.A., Scheidegger, C. & Venkatasubramanian, S. (2021) The (im) possibility of fairness: different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64, 136–143.
Fung, A. (2013) The principle of affected interests: an interpretation and defense. In: Nagel, J.H. & Smith, R.M. (Eds.) *Representation: elections and beyond*. Philadelphia, PA: University of Pennsylvania Press.
Garrett, B.L. & Monahan, J. (2020) Judging risk. *California Law Review*, 108, 439.
Gastwirth, J.L. (1992) Statistical reasoning in the legal setting. *The American Statistician*, 46, 55–69.
Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé, H. et al. (2021) Datasheets for datasets. *Communications of the ACM*, 64, 86–92.
Gelman, A. & Loken, E. (2014) The statistical crisis in science data-dependent analysis—A "garden of forking paths"—Explains why many statistically significant comparisons don't hold up. *American Scientist*, 102, 460.
Gottfredson, D.M. (1987) Prediction and classification in criminal justice decision making. *Crime and Justice*, 9, 1–20.
Grace, Y.Y. (2016) *Statistical analysis with measurement error or misclassification*. New York: Springer.

Green, B. (2020) The false promise of risk assessments: epistemic reform and the limits of fairness. In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. Barcelona, Spain: Association for Computing Machinery, pp. 594–606.
Green, B. & Hu, L. (2018) The myth in the methodology: towards a recontextualization of fairness in machine learning. In: *Proceedings of the machine learning: the debates workshop at ICML*. Stockholm.
Groves, R.M., Fowler, F.J., Jr., Couper, M.P., Lepkowski, J.M., Singer, E. & Tourangeau, R. (2011) *Survey methodology*. New York: John Wiley & Sons.
Habermas, J. (2015) *Between facts and norms: contributions to a discourse theory of law and democracy*. New York: John Wiley & Sons.
Hájek, A. (2007) The reference class problem is your problem too. *Synthese*, 156, 563–585.
Hamilton, M. (2015) Risk-needs assessment: constitutional and ethical challenges. *American Criminal Law Review*, 52, 231.
Hand, D.J. (2006) Classifier technology and the illusion of progress. *Statistical Science*, 21, 1–14.
Hand, D.J. (2009) Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine Learning*, 77, 103–123.
Hannah-Moffat, K. & Shaw, M. (2001) *Taking risks: incorporating gender and culture into the classification and assessment of federally sentenced women in Canada*. Ottawa, ON: Status of Women Canada, Government of Canada.
Hanson, R.K., Babchishin, K.M., Helmus, L.M., Thornton, D. & Phenix, A. (2017) Communicating the results of criterion referenced prediction measures: risk categories for the static-99r and static-2002r sexual offender risk assessment tools. *Psychological Assessment*, 29, 582.
Hardt, M., Price, E. & Srebro, N. (2016) Equality of opportunity in supervised learning. In: *Advances in Neural Information Processing Systems (NIPS)*, Vol. 29. Barcelona, Spain, pp. 3315–3323.
Hart, H. (1961) *The concept of law*. Oxford: Oxford University Press.
Hart, H. & Honoré, T. (1985) *Causation in the law*. Oxford: Oxford University Press.
Hartmann, K. & Wenzelburger, G. (2021) Uncertainty, risk and the use of algorithms in policy decisions: a case study on criminal justice in the USA. *Policy Sciences*, 54, 269–287.
Hastie, T., Tibshirani, R. & Friedman, J. (2009) *The elements of statistical learning: data mining, inference, and prediction*. Cham: Springer.
Hayek, F.A. (1973) *Law, legislation and liberty, volume 1: rules and order*. Chicago: University of Chicago Press.
Heckman, J. (1979) Sample selection bias as a specification error. *Econometrica*, 47, 153–161.
Heinze, G., Wallisch, C. & Dunkler, D. (2018) Variable selection–A review and recommendations for the practicing statistician. *Biometrical Journal*, 60, 431–449.
Hilton, N.Z., Scurich, N. & Helmus, L.-M. (2015) Communicating the risk of violent and offending behavior: review and introduction to this special issue. *Behavioral Sciences & the Law*, 33, 1–18.
Hong, L. & Page, S.E. (2004) Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101, 16385–16389.
Imrey, P.B. & Dawid, A.P. (2015) A commentary on statistical assessment of violence recidivism risk. *Statistics and Public Policy*, 2, 1–18.
Jacobs, A.Z. & Wallach, H. (2021) Measurement and fairness. In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. New York: ACM, pp. 375–385.
Jiang, M. & Fu, K.-W. (2018) Chinese social media and big data: big data, big brother, big profit? *Policy & Internet*, 10, 372–392.
Johnson, K. (2020) *Amsterdam and Helsinki launch algorithm registries to bring transparency to public deployments of AI*. Available from: https://venturebeat.com/2020/09/28/amsterdam-and-helsinki-launch-algorithm-registries-to-bring-transparency-to-public-deployments-of-ai/ [Online, Accessed 1st April 2022].
Jones, M. & Sims, B. (1997) Recidivism of offenders released from prison in North Carolina: a gender comparison. *The Prison Journal*, 77, 335–348.
Kahneman, D., Rosenfield, A., Gandhi, L. & Blaser, T. (2016) Noise. *Harvard Business Review*, 38–46.
Kehl, D.L. & Kessler, S.A. (2017) Algorithms in the criminal justice system: assessing the use of risk assessments in sentencing. In: *Responsive communities initiative, berkman klein center for internet & society*. Cambridge: Harvard Law School.
Kellert, S.H., Longino, H.E. & Waters, C.K. (2006) *Scientific pluralism*. Minneapolis: University of Minnesota Press.

Kiely, T.F. (2005) *Forensic evidence: science and the criminal law*. Boca Raton: CRC Press.
Kitcher, P. (2003) *Science, truth, and democracy*. Oxford: Oxford University Press.
Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. & Mullainathan, S. (2018) Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133, 237–293.
Kleinberg, J., Mullainathan, S. & Raghavan, M. (2016) Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.
Kleinberg, J. & Raghavan, M. (2020) How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)*, 8, 1–23.
Kline, M. (2012) *Mathematics and the physical world*. North Chelmsford: Courier Corporation.
Koepke, J.L. & Robinson, D.G. (2018) Danger ahead: risk assessment and the future of bail reform. *Washington Law Review*, 93, 1725.
Kroll, J.A., Huey, J., Barocas, S., Felten, E.W., Reidenberg, J.R., Robinson, D.G. et al. (2017) Accountable algorithms. *University of Pennsylvania Law Review*, 165, 633–705.
Kuhn, M. & Johnson, K. (2013) *Applied predictive modeling*. New York: Springer.
Kusner, M.J., Loftus, J., Russell, C. & Silva, R. (2017) Counterfactual fairness. In: *Advances in neural information processing systems 30 (NeurIPS)*. Long Beach: Curran Associates, Inc., pp. 4066–4076.
Langan, P. & Levin, D. (2002) *Recidivism of prisoners released in 1994*. Available from: https://www.bjs.gov/content/pub/pdf/rpr94.pdf [Online. Accessed 20th May 2021].
Latour, B. & Woolgar, S. (2013) *Laboratory life*. Princeton: Princeton University Press.
Legomsky, S.H. (2007) Learning to live with unequal justice: asylum and the limits to consistency. *The Stanford Law Review*, 60, 413.
Leiter, B. (2007) *Objectivity in law and morals*. Cambridge: Cambridge University Press.
Li, X. (2020) Research on the building of China's smart court in the internet era. *Chinese Academy of Social Sciences*, 8, 30.
Liu, H.-W., Lin, C.-F. & Chen, Y.-J. (2019) Beyond state v Loomis: artificial intelligence, government algorithmization and accountability. *International Journal of Law and Information Technology*, 27, 122–141.
Lohr, S. (2011) *Sampling: design and analysis*. Boca Raton: CRC Press.
Lovegrove, A. (1997) *The framework of judicial sentencing: a study in legal decision making*. Cambridge: Cambridge University Press.
Luijken, K., Groenwold, R.H., Van Calster, B., Steyerberg, E.W. & van Smeden, M. (2019) Impact of predictor measurement heterogeneity across settings on the performance of prediction models: a measurement error perspective. *Statistics in Medicine*, 38, 3444–3459.
Maltz, M. (2019) *Bridging gaps in police crime data* report. www.bjs.gov/content/pub/pdf/bgpcd.pdf [Accessed 20th May 2021].
Marcus, R.B. (1980) Moral dilemmas and consistency. *The Journal of Philosophy*, 77, 121–136.
Marx, C., Calmon, F. & Ustun, B. (2020) Predictive multiplicity in classification. In: *International conference on machine learning*. PMLR, pp. 6765–6774.
Maxfield, L.D. (2005) Measuring recidivism under the federal sentencing guidelines. *Federal Sentencing Reporter*, 17, 166–170.
Mayson, S.G. (2017) Dangerous defendants. *Yale Law Journal*, 127, 490.
McKay, C. (2020) Predicting risk in criminal procedure: actuarial tools, algorithms, AI and judicial decision-making. *Current Issues in Criminal Justice*, 32, 22–39.
Messick, S. (1995) Validity of psychological assessment: validation of inferences from Persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741.
Miller, T. (2019) Explanation in artificial intelligence: insights from the social sciences. *Artificial Intelligence*, 267, 1–38.
Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B. et al. (2019) Model cards for model reporting. In: *Proceedings of the conference on fairness, accountability, and transparency*. Atlanta, GA. New York, NY: Association for Computing Machinery, pp. 220–229.
Mitchell, S., Potash, E., Barocas, S., D'Amour, A. & Lum, K. (2021) Algorithmic fairness: choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8, 141–163.
Mittelstadt, B. (2019) Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1, 501–507.

Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S. & Floridi, L. (2016) The ethics of algorithms: mapping the debate. *Big Data & Society*, 3, 2053951716679679.

Monahan, J. & Skeem, J.L. (2016) Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology*, 12, 489–513.

Moore, M., Estrich, S., McGillis, D. & Spelman, W. (1984) *Dangerous offenders: the elusive target of justice*. Cambridge: Harvard University Press.

Moses, L.B. & Chan, J. (2014) Using big data for legal and law enforcement decisions: testing the new tools. *University of the New South Wales Law Journal*, 37, 643–678.

Mullainathan, S. & Obermeyer, Z. (2017) Does machine learning automate moral hazard and error? *American Economic Review*, 107, 476–480.

Muller, J.Z. (2019) *The tyranny of metrics*. Princeton: Princeton University Press.

Mulligan, D.K. & Bamberger, K.A. (2019) Procurement as policy: administrative process for machine learning. *Berkeley Technology Law Journal*, 34, 773.

Na, C. & Gottfredson, D.C. (2013) Police officers in schools: effects on school crime and the processing of offending behaviors. *Justice Quarterly*, 30, 619–650.

Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366, 447–453.

Oreskes, N., Shrader-Frechette, K. & Belitz, K. (1994) Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, 263, 641–646.

Oswald, M. (2018) Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376, 20170359.

Partnership on AI. (2016) *Report on algorithmic risk assessment tools in the U. S. criminal justice system*. Available at: www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/ [Online, Accessed 20th May 2021].

Perez, O. (2006) The institutionalization of inconsistency: from fluid concepts to random walk. In: Perez, O. & Taubner, G. (Eds.) *Paradoxes and inconsistencies in the law*. Oxford: Bloomsbury Publishing.

Pettit, B. & Western, B. (2004) Mass imprisonment and the life course: race and class inequality in us incarceration. *American Sociological Review*, 69, 151–169.

Pickering, A. (1995) *The mangle of practice: time, agency, and science*. Chicago, IL: University of Chicago Press.

Pundik, A. (2008) Statistical evidence and individual litigants: a reconsideration of Wasserman's argument from autonomy. *The International Journal of Evidence & Proof*, 12, 303–324.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A. & Lawrence, N.D. (2009) *Dataset shift in machine learning*. Cambridge, MA: The MIT Press.

Ragan-Kelley, B. & Willing, C. (2018) Binder 2.0-reproducible, interactive, sharable environments for science at scale. In: Akici, F., Lippa, D., Niederhut, D. & Pacer, M. (Eds.) *Proceedings of the 17th python in science conference*. Austin, TX, pp. 113–120.

Rawls, J. (2005) *Political liberalism*. New York, NY: Columbia University Press.

Rector, M.G. (1958) Factors in measuring recidivism as presented in annual reports. *Crime & Delinquency*, 4, 218.

Resnik, J. (1982) Managerial judges. *Harvard Law Review*, 96, 374.

Rhee, R.J. (2007) Probability, policy and the problem of reference class. *The International Journal of Evidence & Proof*, 11, 286–291.

Rice, M.E. & Harris, G.T. (1995) Violent recidivism: assessing predictive validity. *Journal of Consulting and Clinical Psychology*, 63, 737.

Ridgeway, G. (2013) The pitfalls of prediction. *NIJ Journal*, 271, 34–40.

Ritter, N. (2013) Predicting recidivism risk: new tool in Philadelphia shows great promise. *National Institute of Justice Journal*, 271, 4–13.

Rosecrance, J. (1988) Maintaining the myth of individualized justice: probation presentence reports. *Justice Quarterly*, 5, 235–256.

Rosenfeld, M. (2000) The rule of law and the legitimacy of constitutional democracy. *Southern California Law Review*, 74, 1307.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L. & Zhong, C. (2022) Interpretable machine learning: fundamental principles and 10 grand challenges. *Statistics Surveys*, 16, 1–85.

Saar-Tsechansky, M. & Provost, F. (2007) Handling missing values when applying classification models. *Journal of Machine Learning Research*, 8, 1623–1657.

Sauerbrei, W., Perperoglou, A., Schmid, M., Abrahamowicz, M., Becher, H., Binder, H. et al. (2020) State of the art in selection of variables and functional forms in multivariable analysis—Outstanding issues. *Diagnostic and Prognostic Research*, 4, 1–18.

Schauer, F. (2009) *Thinking like a lawyer: a new introduction to legal reasoning*. Cambridge, MA: Harvard University Press.

Scherer, M.U. (2015) Regulating artificial intelligence systems: risks, challenges, competencies, and strategies. *The Harvard Journal of Law & Technology*, 29, 353.

Scurich, N. (2018) The case against categorical risk estimates. *Behavioral Sciences & the Law*, 36, 554–564.

Selbst, A.D. & Barocas, S. (2018) The intuitive appeal of explainable machines. *Fordham Law Review*, 87, 1085.

Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S. & Vertesi, J. (2019) Fairness and abstraction in sociotechnical systems. In: *Proceedings of the conference on fairness, accountability, and transparency*. Atlanta, GA, pp. 59–68.

Semenova, L., Rudin, C. & Parr, R. (2019) A study in Rashomon curves and volumes: a new perspective on generalization and model simplicity in machine learning. *arXiv preprint arXiv:1908.01755*.

Shadish, W.R. (1993) Critical multiplism: a research strategy and its attendant tactics. *New Directions for Program Evaluation*, 1993, 13–57.

Shmueli, G. (2010) To explain or to predict? *Statistical Science*, 25, 289–310.

Shmueli, G. (2019) Lift up and act! classifier performance in resource-constrained applications. *arXiv preprint arXiv:1906.03374*.

Shmueli, G., Bruce, P.C., Yahav, I., Patel, N.R. & Lichtendahl, K.C., Jr. (2017) *Data mining for business analytics: concepts, techniques, and applications in R*. New York: John Wiley & Sons.

Simmons, J.P., Nelson, L.D. & Simonsohn, U. (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359–1366.

Simonsohn, U., Simmons, J.P. & Nelson, L.D. (2020) Specification curve analysis. *Nature Human Behaviour*, 4, 1208–1214.

Singh, J.P. (2013) Predictive validity performance indicators in violence risk assessment: a methodological primer. *Behavioral Sciences & the Law*, 31, 8–22.

Singh, J.P., Grann, M. & Fazel, S. (2011) A comparative study of violence risk assessment tools: a systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review*, 31, 499–513.

Slobogin, C. (2017) Principles of risk assessment: sentencing and policing. *Ohio State Journal of Criminal Law*, 15, 583.

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K.P., Singla, A., Weller, A. et al. (2018) A unified approach to quantifying algorithmic unfairness: measuring individual &group unfairness via inequality indices. In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. New York: ACM, pp. 2239–2248.

Stanford Law School Policy Lab. (2019a) *Risk assessment fact sheet*. Available from: https://www-cdn.law.stanford.edu/wp-content/uploads/2019/05/PSA-Sheet-CC-Final-5.10-CC-Upload.pdf/ [Online, Accessed 20th May 2021].

Stanford Law School Policy Lab. (2019b) *Stanford pretrial risk assessment tools factsheet*. Available from: https://law.stanford.edu/pretrial-risk-assessment-tools-factsheet-project/ [Online, Accessed 20th May 2021].

Starr, S.B. (2014) Evidence-based sentencing and the scientific rationalization of discrimination. *The Stanford Law Review*, 66, 803.

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

Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. (2016) Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11, 702–712.

Stevenson, M. (2018) Assessing risk assessment in action. *Minnesota Law Review*, 103, 303.

Sunstein, C., Kahneman, D., Schkade, D. & Ritov, I. (2002) Predictably incoherent judgments. *Stanford Law Review*, 54, 34.

Supreme Court of Iowa. (2017) *Iowa v. gordon*. Available from: https://www.iowacourts.gov/courtcases/754/briefs/1162/embedBrief [Online, Accessed 30th March 2022].

Suresh, H. & Guttag, J.V. (2019) A framework for understanding sources of harm throughout the machine learning life cycle. *arXiv preprint arXiv:1901.10002*.

Tillers, P. (2005) If wishes were horses: discursive comments on attempts to prevent individuals from being unfairly burdened by their reference classes. *Law, Probability and Risk*, 4, 33–49.

Tollenaar, N. & Van Der Heijden, P.G. (2019) Optimizing predictive performance of criminal recidivism models using registration data with binary and survival outcomes. *PLoS One*, 14, e0213245.

Tyler, T.R. (2006) *Why people obey the law*. Princeton: Princeton University Press.

Underwood, B.D. (1979) Law and the crystal ball: predicting behavior with statistical inference and individualized judgment. *The Yale Law Journal*, 88, 1408–1448.

van de Poel, I. (2020) Embedding values in artificial intelligence (AI) systems. *Minds and Machines*, 30, 385–409.

Volokh, E. (2018) Chief justice robots. *Duke LJ*, 68, 1135.

von Hirsch, A. (1984) The ethics of selective incapacitation: observations on the contemporary debate. *Crime & Delinquency*, 30, 175–194.

Wasserman, D.T. (1991) The morality of statistical proof and the risk of mistaken liability. *Cardozo Law Review*, 13, 935.

Weiss, G.M. (2013) Foundations of imbalanced learning. In: *Imbalanced learning: foundations, algorithms, and applications*. Hoboken, NJ: Wiley-IEEE Press, pp. 13–41.

Werth, R. (2019) Risk and punishment: the recent history and uncertain future of actuarial, algorithmic, and εevidence-basedε penal techniques. *Sociology Compass*, 13, e12659.

Whittle, R., Peat, G., Belcher, J., Collins, G.S. & Riley, R.D. (2018) Measurement error and timing of predictor values for multivariable risk prediction models are poorly reported. *Journal of Clinical Epidemiology*, 102, 38–49.

Wolpert, D.H. (1996) The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8, 1341–1390.

Wong, P.-H. (2020) Democratizing algorithmic fairness. *Philosophy & Technology*, 33, 225–244.

Yacoby, Y., Green, B., Griffin, C.L. & Velez, F.D. (2022) εIf it didn't happen, why would i change my decision?ε: how judges respond to counterfactual explanations for the public safety assessment. *arXiv preprint arXiv:2205.05424*.

You, C.-W., Lin, Y.-F., Chuang, Y., Lee, Y.-H., Hsu, P.-Y., Lin, S.-Y. et al. (2018) Sobermotion: leveraging the force of probation officers to reduce the risk of dui recidivism. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2, 1–34.

Završnik, A. (2020) Criminal justice, artificial intelligence systems, and human rights. In: *ERA forum*, Vol. 20. New York: Springer, pp. 567–583.

Završnik, A. (2021) Algorithmic justice: algorithms and big data in criminal justice settings. *European Journal of Criminology*, 18, 623–642.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T. & Dwork, C. (2013) Learning fair representations. In: *International conference on machine learning*. Atlanta, GA: PMLR, pp. 325–333.

Zgoba, K.M. & Dayal, N.P. (2015) Recidivism. *The Encyclopedia of Crime and Punishment*, 1–5.

How to cite this article: Greene, T., Shmueli, G., Fell, J., Lin, C.-F. & Liu, H.-W. (2022) Forks over knives: Predictive inconsistency in criminal justice algorithmic risk assessment tools. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185(Suppl. 2), S692–S723. Available from: https://doi.org/10.1111/rssa.12966