Asian Journal of Criminology
https://doi.org/10.1007/s11417-024-09429-x

Is Machine Learning Really Unsafe and Irresponsible in Social Sciences? Paradoxes and Reconsideration from Recidivism Prediction Tasks

Jianhong Liu[1] · Dianshi Moses Li[1]

Received: 24 May 2024 / Accepted: 25 May 2024

Abstract
The paper addresses some fundamental and hotly debated issues for high-stakes event predictions underpinning the computational approach to social sciences, especially in criminology and criminal justice. We question several prevalent views against machine learning and outline a new paradigm that highlights the promises and promotes the infusion of computational methods and conventional social science approaches.

Keywords Computational social science · Recidivism prediction · Explainable artificial intelligence · Interpretability of algorithms · Black-box algorithms

Introduction

Recent advances and breakthroughs in machine learning, particularly deep learning, have shown superior predictive performance and out-of-sample generalization capabilities (He et al., 2016; Vaswani et al., 2017). This progression has triggered social scientists to ponder the potential contributions of these technologies in their respective fields (Dwyer et al., 2018; Sun et al., 2020). Over the past decade, an array of applications has emerged, leading some to evolve into burgeoning disciplines, including computational social science, computational psychology, and computational education (Kučak et al., 2018; Lazer et al., 2009; Rothenberg et al., 2023).

However, the deployment of these technologies has hit roadblocks in domains such as healthcare and criminal justice, where the opacity of machine learning algorithms has raised concerns due to the high-stakes nature of these domains (Kirchner et al., 2016; Neri et al., 2020). Specifically, the public's trust in these high-stakes arenas—those with profound implications for individuals' or collectives' trajectories—are exceptionally susceptible to the ramifications of erroneous decision-making and obscured decision processes (Rudin et al., 2022; van Dijck, 2022), as erroneous judgments might result in patients

✉ Jianhong Liu
jliu@um.edu.mo

[1] Centre for Empirical Legal Studies, Faculty of Law, University of Macau, Macau, China

Published online: 26 June 2024

receiving inappropriate medical interventions or the inadvertent release of high-stakes criminals, leading to potentially grave consequences (Garrett & Rudin, 2022; Rudin, 2019).

Despite considerable interdisciplinary efforts to integrate machine learning technologies into social science (Corbett-Davies et al., 2023; Simmler et al., 2022), recent scholarship has cast doubt on the foundations of the computational approach to social science. Researchers hailing from distinct academic traditions have grappled with finding a harmonious balance between data-driven methodologies and theory-driven frameworks (Angelino et al., 2017; Dwyer et al., 2018; Rothenberg et al., 2023; Sun et al., 2020). The focus of this multi-disciplinary discourse has been multi-faceted, encompassing the quest for algorithmic accuracy (Ozkan et al., 2020; Singh & Mohapatra, 2021), the imperative for interpretability in algorithmic decision-making (Rudin et al., 2022; Tolan et al., 2019), and the ethical dimensions related to procedural justice (Kaissis et al., 2020; Neri et al., 2020). Nevertheless, emerging literature has not only questioned the efficacy of these cross-disciplinary endeavors but also challenged the very underpinnings of computational approaches to social sciences (Dressel & Farid, 2018; Kirchner et al., 2016; Ozkan et al., 2020; Tolan et al., 2019; Wexler, 2017).

The views against machine learning approaches are primarily on two dimensions.

Firstly, the question of algorithmic accuracy remains a significant hurdle. A thorough systematic review reveals that most studies report only modest performance metrics. The Area Under the Receiver Operating Characteristic (AUC-ROC) scores typically hover around 0.74[1] (Travaini et al., 2022). Dressel and Farid (2018) demonstrated that a simple model using only two features could achieve an accuracy of 68%, comparable to that of an untrained human. Similarly, Etzler et al. (2023) argue that the random forest, often considered a complex black-box, does not outperform logistic regression in recidivism predictions. This skepticism is especially pertinent given that the promise of these algorithms largely hinges on their ability to offer robust predictive performance and generalize well to out-of-sample data (Dwyer et al., 2018; Herm et al., 2023).

Secondly, achieving acceptable levels of interpretability in algorithms remains a formidable challenge. For years, the scientific community has grappled with the trade-off between algorithmic performance and interpretability (Došilović et al., 2018; Gunning & Aha, 2019). Generally, the more interpretable an algorithm, the simpler—and often less accurate—it tends to be (Arrieta et al., 2020; Nanayakkara et al., 2018). The importance of interpretability or explainability in machine learning algorithms is a point of consensus across both computer science and social science literature (Minh et al., 2022; Skeem & Lowenkamp, 2020). This issue is particularly acute in high-stakes applications like recidivism prediction, where interpretability is seen as a crucial countermeasure against potential biases, including racial discrimination and gender inequality (Garrett & Rudin, 2022; Karimi-Haghighi & Castillo, 2021; Rudin et al., 2020a, 2020b; van Dijck, 2022). Consequently, algorithms that lack sufficient interpretability may be considered unsuitable for deployment in such sensitive and high-stakes domains (Rudin, 2019; Wang et al., 2022).

[1] The systematic review indicates a reasonably good accuracy performance with an average score of 81%. However, it is challenging to label the classifier as "good" based solely on accuracy (ACC), as it is not an ideal metric for evaluating classification tasks. The ACC score can be significantly influenced by threshold variations without necessarily improving classification capability. A more robust metric is the area under the receiver operating characteristic (AUC-ROC) score. The score of 0.74 is not excellent for recidivism prediction, as it implies a heightened risk of inaccurate predictions.

Given these formidable difficulties, this paper, however, takes a distinctly different stance from much prevailing literature. By revisiting recent advancements in these high-stakes applications in criminal justice, specifically in recidivism prediction, this paper uncovers a series of inherent paradoxes in prior research. Grounded in these paradoxes, this paper contends that the commonly cited challenges—such as performance limitations, the black-box nature of algorithms, and issues of ethical requirements—are, in fact, significantly *overstated*. The quest for algorithmic interpretability, a concept highly valued by computer scientists, social scientists, and criminologists, should not be considered the ultimate solution to address concerns of machine bias, potential discrimination, and algorithmic irresponsibility, especially in the case of black-box algorithms. Instead, we advocate for a paradigm shift in the research. In this new paradigm, concerns regarding interpretability, accountability, and justice can be addressed holistically by integrating data-driven methods with theory-driven methods. We argue that much of the current criticism of algorithms is misplaced. Addressing these issues requires a collaborative effort among computer scientists, social scientists, and criminologists.

The Myths of Machine Learning's Weak Performance in Recidivism Prediction

Paradox 1: Mistaking the Dated Algorithms for the Advanced

A popular argument is that algorithms do not perform well. However, an intriguing observation emerges upon scrutinizing the landscape of recidivism prediction studies: there is a pronounced reliance on classical machine learning algorithms and statistical methodologies, many of which were proposed several decades ago. To elucidate, according to Travaini et al. (2022), logistic regression (LR), random forest (RF), and support vector machines (SVM) as the most prevalent algorithms in recidivism prediction, were introduced in 1944, 1984, and 1964, respectively (Berkson, 1944; Breiman et al., 1984; Vapnik, 1964). This reliance on dated algorithms raises questions about their contemporary relevance and performance.

Examining the trajectory of fields like computer vision and natural language processing provides a revealing contrast. These domains witnessed revolutionary progress with the integration of deep learning techniques. Advanced architectures like ResNet (He et al., 2016) and Transformer (Vaswani et al., 2017) marked significant inflection points, ushering in unprecedented advancements in AI. These strides were made possible by innovative mechanisms like residual connections (He et al., 2016) and self-attention mechanisms (Vaswani et al., 2017), facilitating the development of sophisticated models adept at challenges such as zero-shot learning and semantic segmentation (Guo et al., 2018; Wang et al., 2019).

However, there seems to be hesitance in the realm of recidivism prediction to incorporate these cutting-edge approaches or technologies based on deep learning backbone, such as attention mechanism and transfer learning, which would significantly enhance out-of-sample generalization (Vaswani et al., 2017; Zhou et al., 2022). The prevalent conclusion that algorithms failed to surpass human beings and challenged all efforts made in algorithmic recidivism prediction is solely based on dated classical algorithms, while overlooking the capabilities of contemporary techniques like deep learning, which seems iniquitous, especially when considering that even in their foundational domains, computer vision and natural language processing, classical algorithms like LR, RF, and SVM did not achieve commendable results until the recent paradigm shifts in deep learning.

Paradox 2: Feeding Machine Learning Algorithms by Data Collected from Traditional Scales and Questionnaires

Historically, in the criminology statistical practice, scholars have primarily employed tabular data for empirical analyses. Such data predominantly stems from official records, victimization surveys, and self-report questionnaires (Carrington et al., 2018; Liu, 2008, 2018). Owing to the constraints of computational power and the desire to avoid "information overload", criminologists have traditionally focused on a limited set of factors, often underpinned by one or a few analytical frameworks (Dwyer et al., 2018; Rothenberg et al., 2023). These practices align with a conventional paradigm that seeks to construct broad theoretical frameworks elucidating social phenomena or individual behaviors. Subsequently, these frameworks generate empirically testable hypotheses that either corroborate or challenge the initial theories—a process emblematic of theory-driven methods (Bronfenbrenner & Morris, 2007; Sun et al., 2020). For instance, within Asian Criminology, scholars might hypothesize that disruptive relations are significant predictors of criminal risk, according to relational approaches (Liu, 2017, 2021, 2022; Liu et al., 2017). This approach is intrinsically tied to the philosophical essence of theory-driven methods, which seek to distill generalizable patterns from empirical data (Liu, 2024).

In stark contrast to theory-driven methods, data-driven approaches are not constrained by the computational and analytic constraints inherent in more traditional statistical techniques. Consequently, data-driven machine learning predictors can encompass a vast array of potential influencing factors, identifying those with the most significant impact (Rothenberg et al., 2023). For instance, automatic content analysis, a machine learning-based data-driven method, is replacing traditional content analysis due to its superior accuracy, cost-effectiveness, and expedited processing (Grimmer & Stewart, 2013; Wankhade et al., 2022). By extracting insights directly from raw content, rather than relying on coders or interviewers as intermediaries—as is customary in traditional content analysis and questionnaires—measurement biases, potential prejudices, and deviations arising from coders' limited perspectives are minimized (Drasgow, 1987; Nelson, 2020; Zhao et al., 2013, 2022). Algorithms excel at processing and modeling the vast range of unstructured data that would overwhelm human capacities (Wankhade et al., 2022). This divergence from the conventional theory-driven paradigm underlines the philosophical underpinnings of data-driven methods, which focus on intricate modeling processes to unveil the nuanced mechanisms concealed within empirical data (Dwyer et al., 2018; Rothenberg et al., 2023). Consequently, it is imperative for data-driven methodologies to be nourished with raw, comprehensive, and intricate data, rather than the data with handcrafted features typically harvested from psychological scales and surveys.

Unfortunately, reviewing data sources typically utilized for recidivism prediction reveals that a majority of such datasets can be categorized as outlined by Liu (2008), namely official records, victimization surveys, and self-reported questionnaires, which can be regarded as tabular datasets. Although some scholars regard such a problem, they just take it as the advantage of simple algorithms (usually more interpretable) in comparison to complex black-box algorithms (Garrett & Rudin, 2023), as simple algorithms perform as well as complex algorithms on tabular data (Garrett & Rudin, 2023; Herm et al., 2023). Such data predominantly come from scales and questionnaires, leading us to our Paradox 2. However, in the realm of social science research, there seems to be no dearth of data suitable for training data-driven machine learning models. Owing to specific historical and cultural contexts, criminology scholars in Asia exhibit a predilection for interview-centric qualitative methodologies as one of the primary methodologies in criminal justice and criminology research—a methodology once deemed antiquated (Carrington et al., 2018; Liu & Wu, 2023; Shuai & Liu, 2023). With the recent advancements in AI analytical tools for unstructured textual and audiovisual data, these interview archives can be transformed into invaluable resources for data-driven machine learning investigations, aiding in the measurement of deeper psychological facets, such as personality traits and charisma (Harrison et al., 2019, 2020), which offer more information for analysis. Furthermore, the emergence of smart prison systems and judicial frameworks within smart city infrastructures (Simmler et al., 2022) promises to provide a wealth of unstructured data, enabling a more comprehensive assessment of an offender's rehabilitation during incarceration and the associated risks post-release (Korobkin & Ulen, 2000).

Paradox 3: Historical Factors Determine Future Human Behavior

Another potential constraint on the efficacy of prediction pertains to the foundational behavioral assumptions underpinning the task. In recent studies, most scholars, particularly those from a computer science background, have trained recidivism prediction algorithms using historical factors (such as age and criminal records) and the binary outcomes of recidivism behavior within a set period (e.g., 6 months or 2 years) (Dressel & Farid, 2018; Etzler et al., 2023; Garrett & Rudin, 2022; Ma et al., 2022; Ozkan et al., 2020; Rudin et al., 2022; Wang et al., 2022). Initially, these historical elements are employed to forecast re-offending. Subsequently, the binary outcome of recidivism is used as a proxy variable for recidivism risk, with some computer scientists also using the model-assigned probability or score as a gauge of an offender's likelihood of re-offense (Etzler et al., 2023; Ma et al., 2022; Wang et al., 2022). While these approaches may seem intuitively compelling, they often reflect an oversimplified and deterministic perspective that contradicts contemporary social science and criminology theories. Historical factors alone are insufficient to predict human actions, a notion that echoes the rationality assumption in classical economics, which sought a formulaic representation of human behavior (Korobkin & Ulen, 2000). This approach is now considered an oversimplification of socio-economic dynamics. Modern theories emphasize the uncertainty and bounded rationality of decision-making, with phenomena such as anchoring, heuristics, and framing suggesting that oversimplified behavioral assumptions can be detrimental to accurately describing human behavior (Kahneman & Tversky, 1972; Tversky & Kahneman, 1981).

Social science and criminological theories provide ample evidence that only historical factors alone cannot predict binary outcomes without additional contextual information, and prediction can only be probabilistic, regardless of which advanced algorithm is employed. An individual's likelihood of re-offending post-release is influenced by a multitude of factors, including personality, background, environment, education, opportunities, social networks, mental health, and unforeseen events (Liu, 2005, 2017, 2021; Messner et al., 2018). These determinants, distinct from historical data, also play a crucial role in influencing re-offense.

In designing predictive models, it is imperative to acknowledge the inherent unpredictability and uncertainty in human behavior concerning both predictive targets and elements. Integrating prior knowledge from social sciences, both in terms of model structure and optimization objectives, can align the algorithm's assumptions more closely with the nuances of real-world socio-systems. It is also vital to recognize that the risk of recidivism evolves with time, environment, and unforeseen factors. Therefore, the risk should not be reduced to a binary outcome nor used as a definitive label for offenders (Wang et al., 2022).

Worship of Interpretability of Algorithms

Paradox 4: Interpretability of Algorithms Produce Justice, Fairness, Non-discrimination, Reliability, Trust by the Public, and Other Ethical Requirements

Almost all the research articles emphasize the importance and necessity of the interpretability of ML algorithms, particularly in high-stakes fields like criminology and criminal justice (Li et al., 2019; Medvedeva et al., 2019; Padovan et al., 2023; Travaini et al., 2022; van Dijck, 2022). An influential publication by Kirchner et al. (2016) from ProPublica criticized the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), a widely-known recidivist risk assessment software, accusing its algorithm of discriminating against Black individuals. ProPublica, an NGO media outlet focused on reporting abuse of power and inequality, published the critique, which sparked a backlash against the use of non-transparent algorithms (Chodosh, 2018; Johndrow & Lum, 2019; Wexler, 2017). Rudin et al., (2020a, 2020b) cited ProPublica's case and noted that, despite statistical flaws in ProPublica's results, it is challenging for ML algorithms to gain public trust unless more interpretable and fair models are proposed. Since then, an increasing number of scholars are asserting the importance of interpretability; it seems become widely accepted that interpretable algorithms provide solutions to the issue of racial and gender discrimination (Johndrow & Lum, 2019; Soares & Angelov, 2019), unfairness (Etzler et al., 2023; Johndrow & Lum, 2019; Rudin et al., 2020a), distrust from public (Rudin et al., 2022; Wang et al., 2022), the need of clarify rights and responsibilities (Rudin et al., 2020b), promoting reliability and validity (Wang et al., 2022), and even enhance the generalization ability on data with different distributions (Rudin et al., 2022).

However, we argue that the concept of interpretability, including its functions, explanation typologies, and the probability of achieving it, is ambiguous (Arrieta et al., 2020; Bathaee, 2017; Castelvecchi, 2016; Herm et al., 2023; Lipton, 2018; Marcinkevics & Vogt, 2020; Mohseni et al., 2021; Rudin, 2019). While the transparency of algorithms is attractive (Ghassemi et al., 2021), it may not be entirely convincing to claim that algorithmic interpretability is the definitive answer to the issues raised.

Nearly all literature implicitly assumes that explainability or interpretability can mitigate potential issues of inequity, injustice, discrimination, and public distrust embedded within algorithms (Medvedeva et al., 2023; Rudin, 2019; Rudin et al., 2020a; Wang et al., 2022). Indeed, some influential authors claim the utmost importance of interpretability:

> To prevent errors, prevent due process violations, allow independent validation of models, and gain public trust, we **must** create interpretable and fair models. (Wang et al., 2022).
>
> This further complicates their adoption in many sensitive disciplines, raising concerns from the ethical, privacy, fairness, and transparency perspectives. The **root cause** of the problem is the lack of explainability and/or interpretability of the decision. (Islam et al., 2020).

In recent works, several studies even posit explainability or interpretability as the sole solution to achieving these aims (Islam et al., 2020; Padovan et al., 2023; Rudin et al., 2020a). Based on these claims, researchers have proposed algorithms for detecting discrimination and metrics for assessing fairness, and indiscriminate behavior prediction models (Green, 2020; Karimi-Haghighi & Castillo, 2021; Wang et al., 2022). Yet, few challenge this widely accepted, almost consensual, premise.

A recent study by Rudin's group presents a paradoxical challenge to these assumptions (Wang et al., 2022). In their study, Wang et al. (2022) employed the Explainable Boosting Machine (EBM) models to predict recidivism, claiming their models are interpretable, fair, and potentially valuable in real-world applications. Their primary approach to elucidate the model's results was through visualization, and by enumerating the importance of each factor influencing recidivism predictions. Notably, the variable "ADE", which denotes the number of times an individual was assigned to alcohol and drug education classes as mandated by Kentucky state law for DUI convictions, was considered more influential than "p_drug", which indicates prior drug-related charges. However, it is likely that an experienced criminologist would regard "p_drug" as indicative of a more severe offense than "ADE". Outcomes that lack theoretical grounding and logical interpretation might confuse public perception rather than enhance understanding.

The interpretation of logistic regression is another contentious issue. Owing to its straightforward structure, logistic regression is often categorized as a white-box model (Freitas, 2019; Zhou et al., 2018). All pertinent information is encapsulated within the coefficients of logistic regression models, and the computational process of the algorithm is wholly transparent. Nonetheless, it is an oversimplification to suggest that logistic regression inherently ensures fairness, justice, and other ethical requirements. Recent studies have warned about the potential risks of logistic regression in high-stakes scenarios. Notably, its coefficients can be disproportionately influenced by imbalanced datasets and sensitive attributes, potentially compromising its fairness (Alikhademi et al., 2021; Dreiseitl & Ohno-Machado, 2002). Even with optimal performance, the fairness of algorithms can be compromised by potential biases lurking in the data. Completely interpretable algorithms alone are insufficient to address the inherent inequalities in the real world, unless someone actively works to advance social structures rather than merely addressing unbalanced data in the lab.

These examples underscore that algorithmic outputs, even from models deemed "interpretable," must be contextualized and critically examined. There is a significant gap between interpretable models and ethical hopes. While the transparency of algorithms largely equates to technical interpretability (Arrieta et al., 2020; Stevens & De Smedt, 2023), it remains at a distinct remove from a genuine human-centric understanding. And even more profound is the gap between this human-centric understanding and the broader, yet crucial, benchmarks of justice, fairness, non-discrimination, reliability, and public trust. Simply achieving transparency is insufficient, and even pairing it with understanding does not necessarily meet the higher ethical and societal standards demanded by the public. Regrettably, a majority of scholars in this domain, particularly those with backgrounds in computer science, often overlook this distinction. They tend to view technical interpretability as the end and target of their research and mistakenly equate explainability, interpretability, or understandability with ethical benchmarks. Their focus on interpretability is predominantly anchored in elucidating models or datasets, rather than unpacking the broader societal phenomena. This oversight is elaborated upon in Paradox 5:

Paradox 5: It Is Enough to Explain or Interpret Algorithm

In recent works, computer scientists and social scientists have considered that explanations or interpretations that are faithful to the original model computed are important (Rudin, 2019; Rudin et al., 2020a). However, there appears to be a significant oversight: the explanations derived from algorithms do not automatically translate to explanations of social phenomena. Meeting the ethical requirements of algorithms largely depends on the explanations or interpretations of social phenomena rather than on the algorithms themselves.

Consider the statement from an article by Rudin and Shaposhnik (2023): “All 500 individuals with a credit history of less than 5 years were predicted to default on a loan”. This algorithm is said to generate faithful explanations, touted to be more useful and safer than black-box models. However, her example of 500 default on loans did not really provide a good example for her purpose, and it does not sufficiently address the critical question: Why are these individuals deemed likely to default on loans? This is akin to noting that 500 patients recovered after taking a mysterious medicine. We remain unaware of the medicine’s workings or its interaction with the ailment; we only know that many have recovered post-administration.

While algorithms of this nature are often touted as inherently interpretable, their interpretations, though faithful to the model and raw data, frequently serve as a new kind of black box. They fail to provide the causal explanations that are essential for understanding the “why” behind their conclusions (Halpern & Pearl, 2005; Pearl et al., 2016). To some extent, even the interpretation of the algorithm remains to contain further black box, as no explanation can completely elucidate all mechanisms in human inquiry. We might term these algorithms “black-box interpretability algorithms”.

Black-box Interpretability Algorithms (BIA) are nearly as unsatisfactory as the original black-box algorithms, while their interpretations may seem meaningful to some extent, they do not resolve the essential problem of black box. Firstly, BIA struggles to address undesirable biases present in the training data. While numerous so-called fairness metrics or benchmarks have attempted to assert their roles in combating inequality (Ma et al., 2022), a systematic analysis has revealed two primary fairness categories: (1) those that limit the decision effects on disparities and (2) those that restrict the decision effects based on legally protected characteristics, such as race and gender. Alarmingly, these two fairness approaches can inadvertently harm the very groups they aim to shield (Corbett-Davies et al., 2023).

Secondly, the pursuit of BIA can obscure the distinction between interpretations related to social phenomena, real-world contexts, and the algorithms themselves. BIA might seem to offer clarity, but it can actually introduce a new layer of complexity. For instance, the Rashomon effect in machine learning, a concept often cited by Rudin, suggests that among a group of models with nearly identical performance, there may be simpler models that provide more interpretable narratives. However, these models, despite similar performance metrics, can offer vastly different interpretations of the same data (Breiman, 2001; Müller et al., 2023; Rudin, 2019). However, if we adhere to the notion that a singular truth exists, the emergence of multiple, seemingly aligned interpretations introduces inherent weaknesses. Interestingly, scholars studying the Rashomon effect posit that none of the multiple interpretations of a single dataset can be confidently declared as the definitive truth (Rudin et al., 2022). Thus, how can we expect interpretability, which may not align with truth, to enlighten us on fairness, justice, and other ethical considerations?

More critically, the pursuit of BIA can overshadow alternative approaches to meeting ethical standards. While proposing a black-box algorithm might invite criticism, both BIA and black-box algorithms fall short in addressing ethical requirements. This can lead to increased operational costs for institutions and diminished engagement with non-interpretable technologies (Bathaee, 2017). Genuine ethical-based interpretability algorithms should guide scholars in understanding social phenomena and delving into the intricate mechanisms of the real world, rather than merely focusing on algorithmic rules. In essence, explainable or interpretable algorithms should offer insights that transcend the algorithms themselves, a notion that leads us to propose Paradox 6.

Paradox 6: Interpretability for Computer Scientists and Engineers Is Same as that for the Criminologists, the Authorities, and the Public

Most systematic reviews converge on the idea that true explainability hinges on providing human-centric, lucid explanations, and emphasizes the primacy of the end users—the human audience—over the mere structural simplicity of the models (Adadi & Berrada, 2018; Arrieta et al., 2020; Islam et al., 2020; Ras et al., 2018; van der Waa et al., 2021; Zhou et al., 2018). However, a critical oversight in these discussions is the question of who exactly constitutes the user. Are they criminologists, regulatory authorities, or the general public? Often, these diverse entities are inadvertently conflated, leading to an oversimplified dialogue that misaligns expectations between algorithm developers and the broader audience (Mohseni et al., 2021).

The challenge lies in the plethora of definitions of explainability stemming from diverse algorithmic origins. The interpretative significance of these definitions varies widely among computer scientists, criminologists, legal professionals, governmental entities, and the general public. Given this diversity, determining which form of explainability to adopt becomes complex, especially when not all interpretations are deemed useful (Ghassemi et al., 2021; Mohseni et al., 2021).

For analytical convenience, consider categorizing users into four primary groups: computer scientists, criminologists, regulatory authorities, and the public. Computer scientists, with their deep technical expertise, may lack a nuanced understanding of sociolegal constructs like fairness, justice, and trust. Criminologists, on the other hand, may be less technologically versed but possess profound insights into these socio-cultural dimensions. Regulatory authorities, being the ultimate end-users of interpretable algorithms, have practical considerations extending beyond mere academic theories (Ghassemi et al., 2021; Mohseni et al., 2021). Lastly, the general public, the end beneficiaries of these models, often lack both technical expertise and specialized sociological training, focusing primarily on their rights and obligations (Miller, 2019; Miller et al., 2017).

Despite the proliferation of explainability techniques, ranging from post hoc rationales to visualizations and local explanations (Arrieta et al., 2020; Li et al., 2022; Stepin et al., 2021), no single method can satisfy the diverse needs of all user groups (Ghassemi et al., 2021; Mohseni et al., 2021). Even transparent models like logistic or linear regression, extensively used across social sciences, are comprehensible predominantly to those with academic training (Brożek et al., 2023). A college undergraduate without mathematical, statistical, or social science training might struggle to interpret such models, and the layperson faces even greater challenges (Batanero et al., 1997; Engel & Sedlmeier, 2011; Zhao et al., 2024).

In essence, each user group has distinct demands for interpretability (Ghassemi et al., 2021; Mohseni et al., 2021). Computer scientists seek explanations to enhance performance and rectify flaws (Amann et al., 2020). Criminologists aim to measure constructs like fairness and justice, abstracting broader social principles from data-driven models. Regulatory bodies focus on accountability, a domain where legal professionals play a critical role. The expectation that the general populace will grasp the intricacies of interpretable algorithms is unrealistic (Miller, 2019; Miller et al., 2017), highlighting the need for tailored approaches to explainability that cater to the unique needs and backgrounds of each user group.

From this perspective, whether the public trusts algorithms fundamentally depends on their confidence in regulatory institutions, which fall within the academic realms of communication and public administration. Authorities should major and elucidate the performance, fairness, and justice of these models, fostering public trust. Instead of hoping for simpler models at the expense of performance, the ideal model, characterized by enhanced performance, fairness, and non-discrimination, should be a collaborative endeavor between computers and criminologists (Miller, 2019).

Paradox 7: Interpretability Is the Sole Way that Leads to Fairness, Justice, Non-discrimination, and Other Ethical Requirements

In recent scholarly discourse, a plethora of literature has emerged advocating for the use of interpretable models as a panacea for achieving fairness, justice, and non-discrimination (Jo et al., 2023; Ma et al., 2022; Rudin, 2019; Rudin et al., 2022; Wang et al., 2022). The assumption that interpretability is synonymous with these lofty ideals is not only prevalent but often portrayed as the key to their realization. However, there exists scant empirical evidence substantiating the claim that interpretability invariably culminates in these desired outcomes (Brożek et al., 2023; Ghassemi et al., 2021; Herm et al., 2023; Miller et al., 2017).

A more nuanced perspective suggests that the pursuit of interpretability invariably entails simplifying or summarizing the behaviors of complex models to render them palatable to human comprehension (Garrett & Rudin, 2022; Wang et al., 2022). This introduces an intricate trade-off: the aspiration to distill models for heightened interpretability might be at loggerheads with the imperative to preserve their predictive prowess (Jo et al., 2023; Minh et al., 2022). Increasing a model's predictive accuracy typically requires increasing its complexity (D'Amour et al., 2022). Human societies, as intricate amalgamations of economic, technological, political, legal, social, historical, cultural, geographical, climatic, and ecological facets, ensure human behaviors are invariably influenced by a mélange of anticipated and unanticipated, deterministic and stochastic variables, exhibiting non-linear and time-variant characteristics (Hong & Wang, 2021, 2023; Liu, 2024). Consequently, hyper-dimensional, hyper-parametric, and profoundly intricate models become indispensable for robust predictions, particularly for out-of-sample forecasting (Dwyer et al., 2018; Yongmiao & Shouyang, 2021).

Given humanity's inherent incapacity to fathom raw data or the labyrinthine intricacies of economic and social machinations, resorting to simplification and abstraction in theoretical constructs (Berk, 2008, 2010), how then can one harbor realistic expectations of representing the extremely complex machinations with a glass-box or white-box model? And then, how can one harbor realistic expectations of completely explicating such a model?

While the primary thrust of this article is not to wade into the philosophical debates surrounding the feasibility of interpretable algorithms, it seeks to recalibrate the misplaced expectations that humanity can solely leverage interpretability techniques to instantiate fairness, justice, trustworthiness, and accountability.

In essence, humanity's tryst with "black boxes" is neither novel nor unprecedented, especially in high-stakes applications.

Throughout scientific history, numerous phenomena have been harnessed and applied long before their complete comprehension materialized. A pertinent analogy in computational social science, or more specifically in the context of recidivism prediction, is the use of medical drugs and devices (Ghassemi et al., 2021). Acetaminophen, known for its analgesic properties, has been a staple in the medical community for over a century. Amid the recent onslaught of COVID-19, acetaminophen retained its stature as one of the most efficacious and safest remedies for alleviating fever symptoms (Galluzzo et al., 2023; Yousefifard et al., 2020). Intriguingly, the mechanistic underpinnings of acetaminophen remain enigmatic, an embodiment of the "black box" paradigm (Kirkpatrick, 2005; Kis et al., 2005). However, this has not deterred regulators, physicians, or patients from placing unwavering trust in its safety profile.

The medical community has navigated the complexities of such "black boxes" through rigorous scientific methodologies like randomized controlled trials (RCTs) (Ghassemi et al., 2021). RCTs assess the efficacy of interventions, be they novel pharmaceuticals, therapeutic regimens, or preventative strategies. Participants are judiciously randomized across experimental groups (subjected to the intervention) and control groups (either receiving a placebo or the standard treatment). Such randomization ensures homogeneity at baseline, thereby mitigating biases. The juxtaposition of the control group allows for an empirical comparison of the intervention's efficacy against either the absence of intervention or alternative interventions. Many RCTs incorporate a "double-blind" design, wherein neither the participants nor the investigators are privy to group allocations, a stratagem that considerably attenuates biases and subjectivity. Upon trial culmination, investigators juxtapose outcomes across groups to infer the intervention's efficacy (Concato et al., 2000; Deaton & Cartwright, 2018). Historically, RCTs have been heralded as the "gold standard" in clinical trials, their design meticulously calibrated to minimize biases and errors, thereby proffering robust empirical evidence even when the intervention remains a "black box" (Bothwell et al., 2016; Torgerson et al., 2015).

Fortuitously, the social sciences offer a wealth of methodological tools analogous to those used in the medical field. Counterfactual causal inference methods provide us with a robust framework for assessing the effects of policies and interventions (Imbens & Rubin, 2015; Rubin, 2005). More precise post hoc inferences fortify our ability to evaluate algorithmic fairness, both from input and output perspectives (Berk, 2020; Pearl, 2009). An array of counterfactual-based causal inference methods endows us with the capacity to assess policy effects rigorously. More precise post-hoc inferences fortify our ability to evaluate algorithmic fairness, both from input and output perspectives (Pearl, 2009). Emerging techniques, such as simulations grounded in authentic datasets, show promise in becoming the de facto "gold standard" for validating predictive tasks within social sciences (Zhao et al., 2022). Drawing upon research in these avant-garde domains, there is a compelling rationale to posit that, analogous to the medical sector, the "black box" of AI in social sciences can also be underpinned by trust and accountability mechanisms. And the first step of this is to acknowledge the existence of BIA, and recognize it is not better than a completely black box. Randomized Controlled Trials (RCTs), quasi-experiments, and post hoc inference methods remain powerful tools, complementary to eXplainable AI (XAI), for scholars and practitioners to explore and validate unknown aspects of AI models. The realization of this vision requires a collaborative effort, bringing together computer scientists, social scientists, and criminologists to forge alternative paths that ensure fairness, trustworthiness, accountability, and other ethical requirements of AI models.

Conclusion

This paper critically examines the application of machine learning algorithms in the field of criminology and criminal justice, particularly focusing on their role in high-stakes decision-making. We challenge prevalent views on performance and interpretability, believing that the complexities of these issues might be overly magnified. The aim of this study is to offer a fresh perspective on integrating machine learning with social science research. We recognize that these critical viewpoints require further empirical evidence for validation. Moreover, it is imperative to propose a new paradigm, one that encompasses a cooperative framework, to further this line of inquiry. To this end, we intend to undertake in-depth empirical research in our upcoming series of articles.

Funding Plano de Investimentos e Despesas de Desenvolvimento da Administração. "Supply and install equipment for the Centre for Empirical Legal Studies, Faculty of Law, University of Macau". Reference number: P-PIDDA-01/FLL/2023. PI: Jianhong Liu.

Data Availability Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Ethical Approval and Informed Consent This article does not contain any studies with human participants performed by any of the authors.

Competing Interests The authors declare no competing interests.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160.
Alikhademi, K., Richardson, B., Drobina, E., & Gilbert, J. E. (2021). Can explainable AI explain unfairness? A framework for evaluating explainable AI. arXiv preprint arXiv:2106.07483.
Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making, 20*(1), 1–9.
Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2017). Learning Certifiably Optimal Rule Lists. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM*.
Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., & Benjamins, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.
Batanero, C., Estepa, A., & Godino, J. D. (1997). Evolution of students' understanding of statistical association in a computer-based teaching environment. In J. B. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics: Proceedings of the 1996 IASE Round Table Conference* (pp. 191–205). Voorburg, The Netherlands: International Statistical Institute.
Bathaee, Y. (2017). The artificial intelligence black box and the failure of intent and causation. *The Harvard Journal of Law & Technology, 31*, 889.

Berk, R. (2010). What you can and can't properly do with regression. *Journal of Quantitative Criminology, 26*, 481–487.
Berk, R. (2008). *Statistical learning from a regression perspective* (Vol. 14). Springer.
Berk, R. (2020). Random Forests. In *Statistical Learning from a Regression Perspective* (pp. 233–295). Springer International Publishing. https://doi.org/10.1007/978-3-030-40189-4_5
Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association, 39*(227), 357–365.
Bothwell, L. E., Greene, J. A., Podolsky, S. H., & Jones, D. S. (2016). *Assessing the gold standard—Lessons from the history of RCTs*. In (Vol. 374, pp. 2175–2181): Mass Medical Soc.
Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199–231.
Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press.
Bronfenbrenner, U., Morris, P. A. (2007). The bioecological model of human development. In W. Damon, R. M., Lerner, R. M. (Eds.), *Handbook of Child Psychology Lerner*. https://doi.org/10.1002/9780470147658.chpsy0114
Brożek, B., Furman, M., Jakubiec, M., & Kucharzyk, B. (2024). The black box problem revisited. Real and imaginary challenges for automated legal decision making. *Artificial Intelligence and Law, 32*(2), 427–440. https://doi.org/10.1007/s10506-023-09356-9
Carrington, K., Hogg, R., Scott, J., & Sozzo, M. (2018). *The Palgrave handbook of criminology and the global south*. Springer.
Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News, 538*(7623), 20.
Chodosh, S. (2018). Courts use algorithms to help determine sentencing, but random people get the same results. Popular Science Available at: https://www.popsci.com/recidivism-algorithm-random-bias. Accessed 18 Jan 2018
Concato, J., Shah, N., & Horwitz, R. I. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine, 342*(25), 1887–1892.
Corbett-Davies, S., Nilforoshan, H., Shroff, R., & Goel, S. (2023). The measure and mismeasure of fairness. *The Journal of Machine Learning Research.*
D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., & Hoffman, M. D. (2022). Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research, 23*(226), 1–61.
Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine, 210*, 2–21.
Došilović, F. K., Brčić, M., & Hlupić, N. (2018). *Explainable artificial intelligence: A survey*. 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO).
Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72*(1), 19.
Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics, 35*(5–6), 352–359.
Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580
Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology, 14*, 91–118.
Engel, J., Sedlmeier, P. (2011). Correlation and regression in the training of teachers. In: Batanero, C., Burrill, G., Reading, C. (Eds.), *Teaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education. New ICMI Study Series, vol. 14*. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-1131-0_25
Etzler, S., Schonbrodt, F. D., Pargent, F., Eher, R., & Rettenberger, M. (2023). Machine learning and risk assessment: Random forest does not outperform logistic regression in the prediction of sexual recidivism. *Assessment*, 10731911231164624. https://doi.org/10.1177/10731911231164624
Freitas, A. A. (2019). Automated machine learning for studying the trade-off between predictive accuracy and interpretability. In Holzinger, A., Kieseberg, P., Tjoa, A., Weippl, E. (Eds.), *Machine Learning and Knowledge Extraction. CD-MAKE 2019*. Lecture Notes in Computer Science, *vol. 11713*. Springer, Cham. https://doi.org/10.1007/978-3-030-29726-8_4
Galluzzo, V., Zazzara, M. B., Ciciarello, F., Tosato, M., Bizzarro, A., Paglionico, A., Varriano, V., Gremese, E., Calvani, R., & Landi, F. (2023). Use of First-Line Oral Analgesics during and after COVID-19: Results from a survey on a sample of Italian 696 COVID-19 survivors with post-acute symptoms. *Journal of Clinical Medicine, 12*(8), 2992.

Garrett, B. L., & Rudin, C. (2022). Glass box artificial intelligence in criminal justice. *Available at SSRN 4275661*.

Garrett, B. L., & Rudin, C. (2023). Interpretable algorithmic forensics. *Proceedings of the National Academy of Sciences, 120*(41), e2301842120.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health, 3*(11), e745–e750.

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 (FAT* '20). Association for Computing Machinery* (pp. 594–606). New York, NY, USA. https://doi.org/10.1145/3351095.3372869

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267–297.

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine, 40*(2), 44–58.

Guo, Y., Liu, Y., Georgiou, T., & Lew, M. S. (2018). A review of semantic segmentation using deep neural networks. *International Journal of Multimedia Information Retrieval, 7*, 87–93.

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science, 56*(4), 843–887.

Harrison, J. S., Thurgood, G. R., Boivie, S., & Pfarrer, M. D. (2019). Measuring CEO personality: Developing, validating, and testing a linguistic tool. *Strategic Management Journal, 40*(8), 1316–1330.

Harrison, J. S., Thurgood, G. R., Boivie, S., & Pfarrer, M. D. (2020). Perception is reality: How CEOs' observed personality influences market perceptions of firm risk and shareholder returns. *Academy of Management Journal, 63*(4), 1166–1195.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Herm, L.-V., Heinrich, K., Wanner, J., & Janiesch, C. (2023). Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. *International Journal of Information Management, 69*, 102538.

Hong, Y., & Wang, S. (2021). How is big data changing economic research paradigms. *Journal of Management World, 37*(10), 40–55.

Hong, Y., & Wang, S. (2023). Impacts of cutting-edge artificial intelligence on economic research paradigm. *Bulletin of Chinese Academy of Sciences (chinese Version), 38*(3), 353–357.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Islam, S. R. , Eberle, W., Ghafoor, S. K. (2020). Towards quantification of explainability in explainable artificial intelligence methods. In Barták, R., Bell, E. (Eds.), *Proceedings of the thirty-third international Florida artificial intelligence research society conference*, originally to be held in North Miami Beach, Florida, USA, May 17–20, 2020. AAAI Press, (pp. 75–81).

Jo, N., Aghaei, S., Benson, J., Gomez, A., & Vayanos, P. (2023). Learning optimal fair decision trees: Trade-offs between interpretability, fairness, and accuracy. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)* (pp. 181–192). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3600211.3604664

Johndrow, J. E., & Lum, K. (2019). An algorithm for removing sensitive information. *The Annals of Applied Statistics, 13*(1), 189–220.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430–454.

Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence, 2*(6), 305–311.

Karimi-Haghighi, M., & Castillo, C. (2021). Enhancing a recidivism prediction tool with machine learning: effectiveness and algorithmic fairness. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law (ICAIL '21)* (pp. 210–214). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3462757.3466150

Kirchner, J., Angwin, S., Mattu, J., & Larson, L. (2016). Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks*. Pro Publica.

Kirkpatrick, P. (2005). New clues in the acetaminophen mystery. *Nature Reviews Drug Discovery, 4*(11), 883–883.

Kis, B., Snipes, J. A., & Busija, D. W. (2005). Acetaminophen and the cyclooxygenase-3 puzzle: Sorting out facts, fictions, and uncertainties. *Journal of Pharmacology and Experimental Therapeutics, 315*(1), 1–7.

Korobkin, R. B., & Ulen, T. S. (2000). Law and behavioral science: Removing the rationality assumption from law and economics. *The California Law Review, 88*, 1051.

Kučak, D., Juričić, V., & Đambić, G. (2018). Machine Learning in Education - a Survey of Current Research Trends. In Katalinic, B. (Ed.), *Proceedings of the 29th DAAAM International Symposium*, (pp. 0406-0410). Published by DAAAM International, Vienna, Austria. https://doi.org/10.2507/29th.daaam.proceedings.059

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., & Gutmann, M. (2009). Computational social science. *Science, 323*(5915), 721–723.

Li, J., Zhang, G., Yu, L., & Meng, T. (2019). Research and design on cognitive computing framework for predicting judicial decisions. *Journal of Signal Processing Systems, 91*, 1159–1167.

Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., Bian, J., & Dou, D. (2022). Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems, 64*(12), 3197–3234.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue, 16*(3), 31–57.

Liu, J. (2005). Predicting recidivism in a communitarian society: China. *International Journal of Offender Therapy and Comparative Criminology, 49*(4), 392–409.

Liu, J. (2008). Data sources in Chinese crime and criminal justice research. *Crime, Law and Social Change, 50*, 131–147.

Liu, J. (2021). Asian criminology and non-Western criminology: Challenges, strategies, and directions. *International Annals of Criminology, 59*(2), 103–118.

Liu, J. (2022). Asian criminology—Elaborating its concepts, approach, paradigm, and future. *Asian Journal of Criminology, 17*(4), 391–399.

Liu, J. (2024). The relationism theory of criminal justice—A paradigm shift. *Asian Journal of Criminology, 19*(1), 1–25. https://doi.org/10.1007/s11417-024-09419-z

Liu, J., Travers, M., & Chang, L. Y. C. (2017). Reflecting on comparison: A view from Asia. *Comparative criminology in Asia*, (pp. 185–201). https://doi.org/10.1007/978-3-319-54942-2_13

Liu J., Wu G. (2023). Procedural fairness and fear of crime: Extending the procedural justice theoretical model under the Chinese context. *Crime & Delinquency.* Advance online publication. https://doi.org/10.1177/00111287221150422

Liu, J. (2017). The New Asian Paradigm: A Relational Approach. In J. Liu, M. Travers, & L. Y. C. Chang (Eds.), *Comparative criminology in Asia* (pp. 17–32). Springer International Publishing. https://doi.org/10.1007/978-3-319-54942-2_2

Liu, J. (2018). The Asian criminological paradigm and how it links Global North and South: Combining an Extended Conceptual Tool box from the North with Innovative Asian Contexts. In Carrington, K., Hogg, R., Scott, J., Sozzo, M. (Eds.), *The Palgrave Handbook of Criminology and the Global South*. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-319-65021-0_4

Ma, Y., Nakamura, K., Lee, E. J., & Bhattacharyya, S. S. (2022). EADTC: An Approach to Interpretable and Accurate Crime Prediction (pp. 170–177). In 2022 *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*.

Marcinkevičs, R., & Vogt, J. E. (2020). *Interpretability and explainability: A machine learning zoo mini-tour*. arXiv preprint arXiv:2012.01805. https://doi.org/10.48550/arXiv.2012.01805

Medvedeva, M., Vols, M., & Wieling, M. (2019). Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law, 28*(2), 237–266. https://doi.org/10.1007/s10506-019-09255-y

Medvedeva, M., Wieling, M., & Vols, M. (2023). Rethinking the field of automatic prediction of court decisions. *Artificial Intelligence and Law, 31*(1), 195–212.

Messner, S. F., Liu, J., & Zhao, Y. (2018). Predicting re-incarceration status of prisoners in contemporary China: Applying Western criminological theories. *International Journal of Offender Therapy and Comparative Criminology, 62*(4), 1018–1042.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. arXiv preprint arXiv:1712.00547.

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review, 55*(5), 3503–3568. https://doi.org/10.1007/s10462-021-10088-y

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS), 11*(3–4), 1–45.
Müller, S., Toborek, V., Beckh, K., Jakobs, M., Bauckhage, C., Welke, P. (2023). An empirical evaluation of the rashomon effect in explainable machine learning. In Koutra, D., Plant, C., Gomez Rodriguez, M., Baralis, E., Bonchi, F. (Eds.), *Machine Learning and Knowledge Discovery in Databases: Research Track*. ECML PKDD 2023. Lecture Notes in Computer Science, vol. 14171. Springer, Cham. https://doi.org/10.1007/978-3-031-43418-1_28
Nanayakkara, S., Fogarty, S., Tremeer, M., Ross, K., Richards, B., Bergmeir, C., Xu, S., Stub, D., Smith, K., & Tacey, M. (2018). Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study. *PLoS Medicine, 15*(11), e1002709.
Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research, 49*(1), 3–42.
Neri, E., Coppola, F., Miele, V., Bibbolino, C., & Grassi, R. (2020). *Artificial intelligence: Who is responsible for the diagnosis?* In (Vol. 125, pp. 517–521): Springer.
Ozkan, T., Clipper, S. J., Piquero, A. R., Baglivio, M., & Wolff, K. (2020). Predicting sexual recidivism. *Sexual Abuse, 32*(4), 375–399.
Padovan, P. H., Martins, C. M., & Reed, C. (2023). Black is the new orange: How to determine AI liability. *Artificial Intelligence and Law, 31*(1), 133–167.
Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.
Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys, 3*, 96–146. https://doi.org/10.1214/09-Ss057
Ras, G., van Gerven, M., Haselager, P. (2018). Explanation methods in deep learning: Users, values, concerns and challenges. In Escalante, H., et al (Eds.), *Explainable and Interpretable Models in Computer Vision and Machine Learning*. The Springer Series on Challenges in Machine Learning. Springer, Cham. https://doi.org/10.1007/978-3-319-98131-4_2
Rothenberg, W. A., Bizzego, A., Esposito, G., Lansford, J. E., Al-Hassan, S. M., Bacchini, D., Bornstein, M. H., Chang, L., Deater-Deckard, K., Di Giunta, L., Dodge, K. A., Gurdal, S., Liu, Q., Long, Q., Oburu, P., Pastorelli, C., Skinner, A. T., Sorbring, E., Tapanya, S., & Alampay, L. P. (2023). Predicting Adolescent Mental Health Outcomes Across Cultures: A Machine Learning Approach. *Journal of Youth and Adolescence, 52*(8), 1595–1619. https://doi.org/10.1007/s10964-023-01767-w
Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100*(469), 322–331.
Rudin, C., & Shaposhnik, Y. (2023). Globally-consistent rule-based summary-explanations for machine learning models: Application to credit-risk evaluation. *Journal of Machine Learning Research, 24*(16), 1–44.
Rudin, C., Wang, C., & Coker, B. (2020a). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review, 2*(1), 1.
Rudin, C., Wang, C., & Coker, B. (2020b). Broader issues surrounding model transparency in criminal Justice Risk Scoring. *Harvard Data Science Review, 2*(1).
Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys, 16*, 1–85.
Rudin, C. (2019). Stop Explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
Shuai, H., & Liu, J. (2023). The relationship between criminology and criminal law: Implications for developing Chinese criminology. *Humanities and Social Sciences Communications, 10*(1), 1–11.
Simmler, M., Brunner, S., Canova, G., & Schedler, K. (2022). Smart criminal justice: Exploring the use of algorithms in the Swiss criminal justice system. *Artificial Intelligence and Law, 31*(2), 213–237. https://doi.org/10.1007/s10506-022-09310-1
Singh, A., & Mohapatra, S. (2021). Development of risk assessment framework for first time offenders using ensemble learning. *IEEE Access, 9*, 135024–135033.
Skeem, J., & Lowenkamp, C. (2020). Using algorithms to address trade-offs inherent in predicting recidivism. *Behavioral Sciences & the Law, 38*(3), 259–278.
Soares, E., & Angelov, P. (2019). Fair-by-design explainable models for prediction of recidivism. arXiv preprint arXiv:1910.02043. https://doi.org/10.48550/arXiv.1910.02043
Stepin, I., Alonso, J. M., Catala, A., & Pereira-Fariña, M. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access, 9*, 11974–12001.

Stevens, A., & De Smedt, J. (2023). Explainability in process outcome prediction: Guidelines to obtain interpretable and faithful models. *European Journal of Operational Research, 317*(2), 317–327. https://doi.org/10.1016/j.ejor.2023.09.010

Sun, X., Ram, N., & McHale, S. M. (2020). Adolescent family experiences predict young adult educational attainment: A data-based cross-study synthesis with machine learning. *Journal of Child and Family Studies, 29*, 2770–2785.

Tolan, S., Miron, M., Gómez, E., & Castillo, C. (2019). Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in catalonia. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law (ICAIL '19)* (pp. 83–92). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3322640.3326705

Torgerson, C.J., Torgerson, D.J. & Taylor, C.A. (2015). Randomized Controlled Trials. In Newcomer, K. E., Hatry, H. P., & Wholey, J. S. (Eds.), *Handbook of Practical Program Evaluation*. https://doi.org/10.1002/9781119171386.ch7

Travaini, G. V., Pacchioni, F., Bellumore, S., Bosia, M., & De Micco, F. (2022). Machine learning and criminal justice: A systematic review of advanced methodology for recidivism risk prediction. *International Journal of Environmental Research and Public Health, 19*(17), 10594. https://doi.org/10.3390/ijerph191710594

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453–458.

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence, 291*, 103404.

van Dijck, G. (2022). Predicting recidivism risk meets AI Act. *European Journal on Criminal Policy and Research, 28*(3), 407–423.

Vapnik, V. N. (1964). A note on one class of perceptrons. *Automat. Rem. Control, 25*, 821–837.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

Wang, W., Zheng, V. W., Yu, H., & Miao, C. (2019). A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST), 10*(2), 1–37.

Wang, C., Han, B., Patel, B., & Rudin, C. (2023). In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction. *Journal of Quantitative Criminology, 39*(2), 519–581. https://doi.org/10.1007/s10940-022-09545-w

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review, 55*(7), 5731–5780.

Wexler, R. (2017). When a computer program keeps you in jail. *The New York Times, 13*, 1.

Yongmiao, H., & Shouyang, W. (2021). Big data, machine learning and Statistics: Challenges and opportunities. *China Journal of Econometrics, 1*(1), 17.

Yousefifard, M., Zali, A., Zarghi, A., Madani Neishaboori, A., Hosseini, M., & Safari, S. (2020). Non-steroidal anti-inflammatory drugs in management of COVID-19; A systematic review on current evidence. *International Journal of Clinical Practice, 74*(9), e13557.

Zhao, X., Liu, J. S., & Deng, K. (2013). Assumptions behind intercoder reliability indices. *Annals of the International Communication Association, 36*(1), 419–480.

Zhao, X., Feng, G. C., Ao, S. H., & Liu, P. L. (2022). Interrater reliability estimators tested against true interrater reliabilities. *BMC Medical Research Methodology, 22*(1), 232.

Zhao, X., Li, D. M., Lai, Z. Z., Liu, P. L., Ao, S. H., & You, F. (2024). Percentage Coefficient (bp)--effect size analysis (Theory Paper 1). arXiv preprint arXiv:2404.19495.

Zhou, Q., Liao, F., Mou, C., & Wang, P. (2018). Measuring interpretability for different types of machine learning models. In M. Ganji, L. Rashidi, B. C. M. Fung, & C. Wang (Eds.), *Trends and Applications in Knowledge Discovery and Data Mining Cham*.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. (2023). Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(4), 4396–4415. https://doi.org/10.1109/TPAMI.2022.3195549