information

Review

Ethics and Trustworthiness of AI for Predicting the Risk of Recidivism: A Systematic Literature Review

Michael Mayowa Farayola [1,*], Irina Tal [1], Regina Connolly [2], Takfarinas Saber [3] and Malika Bendechache [3]

[1] School of Computing, Dublin City University, D09 DXA0 Dublin, Ireland; irina.tal@dcu.ie
[2] School of Business, Dublin City University, D09 DXA0 Dublin, Ireland; regina.connolly@dcu.ie
[3] School of Computer Science, University of Galway, H91 TK33 Galway, Ireland; takfarinas.saber@nuigalway.ie (T.S.); malika.bendechache@universityofgalway.ie (M.B.)
* Correspondence: michael.farayola2@mail.dcu.ie

Abstract: Artificial Intelligence (AI) can be very beneficial in the criminal justice system for predicting the risk of recidivism. AI provides unrivalled high computing power, speed, and accuracy; all harnessed to strengthen the efficiency in predicting convicted individuals who may be on the verge of recommitting a crime. The application of AI models for predicting recidivism has brought positive effects by minimizing the possible re-occurrence of crime. However, the question remains of whether criminal justice system stakeholders can trust AI systems regarding fairness, transparency, privacy and data protection, consistency, societal well-being, and accountability when predicting convicted individuals' possible risk of recidivism. These are all requirements for a trustworthy AI. This paper conducted a systematic literature review examining trust and the different requirements for trustworthy AI applied to predicting the risks of recidivism. Based on this review, we identified current challenges and future directions regarding applying AI models to predict the risk of recidivism. In addition, this paper provides a comprehensive framework of trustworthy AI for predicting the risk of recidivism.

Keywords: trustworthy AI; criminal justice system; trust; recidivism; privacy and data protection

check for updates

Citation: Farayola, M.M.; Tal, I.; Connolly, R.; Saber, T.; Bendechache, M. Ethics and Trustworthiness of AI for Predicting the Risk of Recidivism: A Systematic Literature Review. *Information* **2023**, *14*, 426. https://doi.org/10.3390/info14080426

Academic Editor: Maanak Gupta

Received: 31 May 2023
Revised: 4 July 2023
Accepted: 19 July 2023
Published: 27 July 2023

1. Introduction

Artificial Intelligence (AI) has been a part of the criminal justice system for many years [1]. Criminal justice managers use AI to predict the risk of recidivism, that is, whether a convicted individual will recommit a crime. The prediction can be based on several factors associated with the criminal, such as educational background or previous employment [2]. These AI systems, mainly called risk assessment tools, are beneficial for forecasting and reducing incarceration rates and racial disparities [3,4]. Furthermore, these risk assessment tools enhance decision-making in the criminal justice system and increase public safety [5]. Risk assessment tools are pretrial and post-conviction risk assessment tools. The former focuses on offenders failing to appear in court and the latter focuses on the long-term recidivism of offenders after a release on parole or probation [6,7]. For this literature review, we focused on post-conviction risk assessment. A commonly used assessment tool, particularly in the United States, is Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). COMPAS is used to assess the potential recidivism risk of offenders. The United States of America has the highest rate of incarceration in the world, an issue that has worsened over recent decades [8]. Before the inception of algorithmic predictive tools, recidivism prediction had been at the discretion and intuition of criminal justice officers or based on statistical calculation. However, because of human biases and the reformation needed in the criminal justice system, a more effective assessment of recidivism risks was encouraged in the criminal justice system.

The reformation and mitigation of human bias led to adopting an AI model called risk assessment tools [7,9,10]. These tools have predicted the possible risk of recidivism of more than a million offenders in the United States. However, criminal justice system stakeholders have questioned and scrutinized the recidivism risk assessment tools in the past few years due to the ethical impact on convicted individuals and their trustworthiness [11–13]. In particular, fairness, transparency, privacy, and accountability have been the subjects of concern for stakeholders [14]. One of the criticisms of the predictive assessment systems is that they should be limited in the criminal justice system because of their probabilistic estimations of humans and pattern understanding tendencies rather than the underlying criminal causes. That is, predictive algorithms cannot understand the realistic situations of things [15].

In May 2016, ProPublica investigated COMPAS and found racial disparities in the predictive decision of the system, with the false-positive rates nearly twice as high for black offenders compared to white offenders [16]. The COMPAS analysis opened many critics' eyes to the adverse effects of risk assessment tools for predicting recidivism risk. In addition, COMPAS was tagged as a sexist algorithm because it shows a disparate impact against the female gender [17]. Although it is a known fact that the female gender has a lower recidivism rate than men, the question is, does the COMPAS system consider that in its development? However, recent literature [18,19] has refuted the ProPublica analysis of COMPAS, claiming that ProPublica's procedures were statistically wrong due to their classification definition of terms, the use of model errors, and lack of an analytical approach to showing that predictive algorithms can make predictions free of racial bias. Nevertheless, these discussions and issues challenge the credibility of COMPAS and the application of risk assessment tools for predicting recidivism risks in the criminal justice system.

In addition, privacy and data protection are major priorities for criminal justice systems stakeholders. It is important to note that offenders' personal information forms the dataset used to develop risk assessment tools for predicting recidivism risks; however, to what extent do these risk assessment tools comply with ethical principles and have due regard for fundamental rights (privacy) of offenders and to what extent are their data protected? These questions surface because of the increasing amount of offenders' information that needs protection and AI models for predicting recidivism risks that are susceptible to cyberattacks [2]. There is a need to balance the potential benefits of the recidivism risk assessment tools against private-life interests.

For these risk assessment models to gain public trust in predicting the risk of recidivism of convicted offenders, there is a need to create a framework that will necessitate trustworthy AI models [4,20]. One of the most important factors influencing this research is that the predictive models for recidivism predictions lack privacy and data protection. In addition, these risk assessment tools are unfair [21,22], with fairness being a social and ethical concept and not only a statistical one [23,24]. A must-tackle problem is how criminal justice system stakeholders can trust risk assessment tools to predict the risk of convicted offenders recommitting a crime when released on probation, parole, or after serving time in jail. Therefore, this paper looked at the systematic literature review on "Ethics and Trustworthiness of AI in Predicting the Risk of Recidivism in the Criminal Justice System". We focused on trustworthy AI because it leverages different essential requirements, including privacy and data protection, that will address the various concerns of criminal justice system stakeholders. Lastly, we extended the proposed framework of trustworthy AI by the European Commission.

Several studies relating to trustworthy AI for predicting recidivism risk focused on one or more aspects of trustworthy AI. Several other studies focused on the theoretical concept of trustworthy AI, such as in [25–33], or focused on other domains (e.g., robotics, healthcare etc.). These theoretical reviews are discussed in detail in the related works section.

This paper is structured as follows: background study, related works, methodology, proposed extended requirements for trustworthy AI in the criminal judiciary system, extended analysis of trustworthy AI requirements, issues and challenges associated with the application of AI in the judiciary system, and conclusion and future works.

2. Background Study

The following section discusses the fundamental concepts of trust, trustworthy AI, and the requirements for a trustworthy AI.

2.1. Trust

Trust is a crucial component of success for risk assessment tools for recidivism to thrive [11,15]. Trust is a complex area of focus that has drawn the attention of many practitioners and scholars across different disciplinary fields [34]. This attention has focused on understanding the antecedents of trust, the conceptualization of trust, forms or types of trust, people involved in trust, and how trust impacts our ethics and the associations at different levels of life. Despite this, there is a surprising lack of consensus regarding how trust should be defined. Nevertheless, we provide some definitions and fundamental concepts of trust across several domains, as it will serve as a prerequisite to what trustworthy AI entails.

One overview of trust in commercial and personal transactions in the digital age [35] describes trust as an interpersonal phenomenon that facilitates human relationships by reducing uncertainty risks. Trust is the confidence level a trustor has in a trustee to do the right things. Trust can be attitudinal, predictable, and voluntarist. The attitudinal view of trust focuses on a trusting attitude due to personal beliefs. The predictability view of trust focuses on the notion of the positive expectation of a trustor, based on the trustee's behavior, that the trustee will act benevolently. Lastly, the voluntarist view of trust is the state of voluntary subjection of a trustor to the vulnerability of a trustee, that is, a position of risk. Therefore, we can affirmatively say that trusting involves risk. A study on public trust in local government, explaining the role of good governance practices [36], defines trust as a psychological state that constitutes a willingness to take risks based upon positive expectations of a trustee's behavior. Therefore, trust is a bridge between a trustor and a trustee and a lubricating factor for a consistent relationship. However, to strengthen trust, there is a need for transparency, accountability, and responsiveness on the part of both parties [37].

More on the definition of trust, a discussion on the significance of trust for organizational accountability [38] defined trust as holding a trustee accountable over time with the notion that they will exhibit integrity and honesty. Trust is conceptualized as mutual expectations and reciprocity between the trustor and trustee, strengthening social interaction. Trust also involves the reduction of complexity, ethical accountability, and responsibility. Furthermore, trust between different parties is instituted based on standard norms. In essence, trust can act as a sense of accountability on the part of a trustee and the state of a trustor's vulnerability. In conclusion, a review on trustworthy AI [39] defined trust as the willingness of a trustor to depend on a trustee due to a lack of control over the trustee, thereby making available the opportunistic behavior of the trustee.

It is worth noting that trust develops over time based on the trustee's behavior and conformation to a trustor's beliefs. Therefore, it is essential to understand what precedes the establishment of trust, referred to as the antecedent of trust [32,35,36,40]. The antecedents of trust are ability, benevolence, integrity, and predictability. Ability refers to a trustee's skills, characteristics, and level of competence in a specific domain. Benevolence is a trustee's impulse willingness to do good to a trustor, putting aside self-gain profit. Integrity is the perception that the trustee will always act with consistent and positive values. Predictability assures trust will be sustained throughout the relationship between parties. In all, ability, benevolence, integrity, and predictability can be bracketed as attributes used in judging the trustworthiness of a trustee by a trustor.

Now that we have established the basic concept of trust, the question is, what is the relation between people, trust, and technology, such as AI systems? The relationship is such that people and societies are the trustors, the AI system is the trustee, and the connecting bridge is the trust. For people and societies to trust AI systems and subject themselves to a position of risk and vulnerability, the system needs to be trustworthy. Another question now is what trustworthiness is. Trustworthiness is being competent and committed to doing and

achieving the expectation of a trustor, and trustworthiness is regarded as a virtue possessed by a trusted party [41]. The guidelines set up by the European Commission (EC) [42] defined trustworthiness as a prerequisite need for stakeholders(i.e., people affected by AI systems) to develop, deploy and use AI systems. This scenario led to what is now commonly called Trustworthy AI.

Before diving into what trustworthy AI entails, there is a need to understand other qualities that influence stakeholders' trust in AI systems. There are three qualities: human, environmental, and technological qualities [32]. Human qualities or attributes are associated with unique cultural backgrounds, past experiences, and ideologies. These qualities determine the extent to which an individual will voluntarily be subject to a state of risk at the expense of the trustee's freedom. Environmental qualities entail elements that propagate the level of trust in the deployed environment of AI technologies. These elements include the environment's cultural background, educational system, environmental awareness, technological advancement level, and technology tasks. Technical qualities focus on efficiency in yielding results, conformation to a level of expected performance, and processes in achieving outcomes. In [39], trust in technology is further classified based on the technology functionality, helpfulness, reliability, predictability, performance, purpose, and process.

2.2. Trustworthy AI

The benefits of AI in different spheres of life cannot be overemphasized. However, different conditions necessitate AI systems to be considered trustworthy. Several issues are related to developing and deploying AI models, such as violating individual privacy, racial bias, misunderstanding of its processes, and decision-making.

Trustworthy AI encapsulates the must-have qualities of the AI that warrant ethical approaches [32]. A review of trustworthy AI states that incorporating trust in AI's development and design will enable stakeholders to fully realize its potential [39,42]. A study [43] defined trustworthy AI as fair, secure, robust, transparent, safe, and explainable systems regarding human privacy and fundamental rights, and stakeholders involved in its development, deployment, and use are accountable.

In 2019, the European Commission (EC) developed "The Ethics Guideline for Trustworthy AI [42,44]". According to EC guidelines, trustworthy AI should be ethical, lawful, and robust, creating a foundation for stakeholders to trust AI systems' development, deployment, and usage. The guidelines provided four Ethical principles (i.e., respect for human autonomy, prevention of harm, fairness, and explicability) and a list of seven requirements for trustworthy AI (i.e., human agency and oversight, technical robustness and safety, privacy and data governance, transparency, non-discrimination and fairness, societal and environmental wellbeing, and accountability). A point to note is that these seven requirements are non-exhaustive, meaning several other requirements still apply to different domains. Still, these seven can serve as the base for any public or private sector considering trustworthy AI in its activities [42].

This paper conducted the first systematic literature review of AI's ethics and trustworthiness in predicting recidivism risk. To address and assess the requirements, trustworthy AI can be evaluated through technical and non-technical methods. Our research scope looked at the technical approaches to achieving a trustworthy AI more in-depth, which serve as a future research direction. These technical methods will cut across architectures for trustworthy AI, ethics, and the rule of law by design, explanation methods, testing and validating, and quality of service indicators. In this paper, looking at the seven requirements of a trustworthy AI proposed by the EC was essential. These requirements serve as a baseline for our extended requirements for trustworthy AI in predicting the risk of recidivism. The discussion of these requirements is found in Section 5.

2.3. An overview of the Seven Requirements for Trustworthy AI Proposed by the European Commission

The following discussed requirements of trustworthy AI are the requirements proposed by the EC that can be applied and serve as the basis for any public or private domain.

For every field, other requirements need to be considered and added to the proposed seven requirements by the European Commission to actualize trustworthy AI in such fields.

2.3.1. Human Agency and Oversight

Human agency and oversight revolve around fundamental human rights, human agencies, and the human administration of AI systems. AI systems should be built to support human autonomy in decision-making and not infringe on their fundamental rights. AI users should be at liberty to make decisions without AI systems making an adverse impact. This act will give AI users a sense of responsibility and freedom and enable trust in AI technology. AI systems should provide the required information to their users to better understand and interact with the system, allowing users to challenge the AI system's decisions when needs arise. Lastly, humans should engage in the decision process (human-in-the-loop), design cycle (human-on-the-loop), and the overall activities of the AI system (human in command).

2.3.2. Technical Robustness and Safety

AI systems are beneficial to the human race. However, if proper mechanisms are not in place, AI systems can cause harm. AI systems should bring safety to their users and prevent harm in every possible instance. AI systems contain data information, and an attack on the system can influence its outcomes leading to biases or harm to society. AI systems must be secure and built to withstand external attacks or threats. In adverse situations, AI systems should have a fallback plan to safeguard users and data information.

2.3.3. Diversity, Non-Discrimination, Fairness

Fairness is a requirement that has received the focus of many AI stakeholders since the advent of AI systems. Fairness must constantly be taken into account while developing AI systems. AI systems are vulnerable to prejudice if the AI development design lacks appropriate bias mitigation techniques. Hence, the developers should ensure the development of the AI system is void of discrimination and bias. In conclusion, AI systems should be inclusive and accessible to all social groups irrespective of their demographic information.

2.3.4. Accountability

Accountability is a requirement that enables the trust of AI stakeholders in AI systems when there is a level of responsibility and answerability. AI systems are inanimate tools, interacting with humans and influencing the decisions of their users directly or indirectly. Organizations should bear full responsibility in cases of negative impact caused by AI systems at different user instances. Users trust the system more when there is a level of answerability for its decision-making, especially with adverse effects.

2.3.5. Transparency

Transparency entails deliberate documentation, detailing, and understanding of AI systems, such as the data collection processes, design processes, and purpose of building such AI systems. When understanding the system's underlying structure, transparency enables smooth auditing of AI systems. In essence, the procedures followed throughout the AI system design should be well-documented and answerable for issues related to the AI system. Transparency gives a head start on why AI systems behave in a particular manner and produce its outcome.

2.3.6. Privacy and Data Governance

Data are crucial in the development of AI systems. Apart from the models used, AI systems mainly function based on the consumption of large datasets used in designing the system. It is unarguably vital for developers to put in efforts toward the quality of data used for developing AI systems. Data are a significant source of biases in AI systems. Data tend to be biased without mitigating procedures to curtail bias. In addition, data privacy is

paramount in the development of trustworthy AI. Access to the data should be restricted to authorized personnel only.

2.3.7. Societal and Environmental Wellbeing

AI systems have numerous benefits and have come to stay. Developers must build AI systems in such a way that they do not cause harm to humans. They should be designed to enhance humans' capabilities and not impose on their fundamental rights. AI systems should be eco-friendly, sustainable, and maintained. Lastly, developers and authorized stakeholders should oversee the AI system at all times to detect possible adverse effects it may cause to its users and the environment.

The following section discusses the existing works on the requirements for a trustworthy AI.

3. Related Works

This section discusses related works that review or survey the requirements for building a trustworthy AI to predict recidivism risks.

One of the essential propelling factors for AI existence is datasets [45]. A challenging factor for AI systems when predicting recidivism risk is the problem of biases [22,46,47]. Different biases influence the predictive outcomes of AI algorithms that predict recidivism risk. These biases are found in the sample representatives, label features, feature engineering, modelling pipeline, and program implementation [16,48]. Examples of these biases' sources stem from the offender's race, ethnicity, and gender [9,48,49]. A survey on the accuracy and fairness of juvenile justice risk assessment [50] stated that there is always the presence of racial bias in the dataset used in training risk assessment tools for predicting the risk of recidivism. This racial bias often leads to unequal treatment, opportunity, or outcome for convicted individuals [51]. A systematic bias embedded in the dataset used for training predictive risk assessment models can result in violations of offenders' rights [4]. Research has affirmed that these tools are unfair and can be biased against a group of individuals [7,22,48,52–55]. Biases found in risk assessment tools are mostly into three layers: data layer, model layer, and evaluation layer. However, many scholars over-concentrate on one of the layers instead of considering all layers simultaneously. For example, a few pieces of literature address the model layer in recidivism risk assessment tools while neglecting the other layers, and those addressing other layers focus only on the data or evaluation layer [47]. Therefore, a future research direction to ensure trustworthy AI for predicting recidivism risk is to address bias at all layers simultaneously.

The current literature on risk assessment tools has no consensus definition of fairness when predicting offenders' recidivism risk [11,56]. Today's available definitions of fairness in the literature are not optimized enough to affirm what fairness in predicting recidivism should be. Nevertheless, scholars are still debating on reaching a generally acceptable definition of fairness for risk assessment tools for recidivism prediction. A review of fairness in criminal justice risk assessments [56] stated that until scholars precisely define fairness when predicting recidivism risk, it is crucial to do away with indefensible claims of definitions of fairness before being implemented in an unruly manner into policy. It is worth noting that, despite the various definitions of fairness applicable to risk assessment tools for predicting recidivism, there are trade-offs amidst the definitions. Therefore, there is a dire need for a generally acceptable definition of fairness when predicting the risk of recidivism.

Understanding the human perception of defining a generally acceptable definition of fairness when predicting the risk of recidivism is crucial. A survey in [57] gave a comparative approach that could help understand how fair decisions should be defined. The approach centred around eight latent properties (reliability, relevance, privacy, volitionality, causes outcome, causes vicious cycle, causes disparity in outcomes, caused by sensitive group membership) and a given question: is it fair to use a feature in a shared decision-making scenario? This survey helps us to better understand whether the presence of features in describing an offender impacts fairness. From the study in [57], the conclusion made is that (1) the presence of features in defining an offender does not necessarily

lead to discrimination against the offender, (2) there is a lack of a common ground on which features to be considered fair or unfair when developing risk assessment tools for recidivism, (3) people tend to have common reasoning when making a judgment based on the latent properties, and (4) six of the latent properties are statistically necessary for ascertaining fairness judgments.

Transparency is essential when building trust and developing risk assessment tools for recidivism prediction [8,10,13]. Transparency aid the citizens' understanding of the systems' impact and questioning of the system. A report from England and Wales found a need for explicit AI systems' transparency when applied to the criminal justice system [58]. Additionally, the need for transparency will facilitate fairness and accountability by ensuring that companies and governments are aware of the impact of risk assessment tools when predicting the risk of recidivism. However, there are contentions against full transparency of these risk assessment tools for predicting recidivism because of the possibilities of (i) leakage of sensitive data to the public; (ii) backfiring into an implicit invitation to game the system; (iii) a direct impact on the company's competitiveness and developer's reputation; and (iv) inherent opacity of algorithms, whose interpretability may be hard for experts [8]. All these arguments limit the full transparency of recidivism risk assessment algorithms.

The private and commercial sectors own risk assessment tools for predicting recidivism risk. The problem with these sectors is that they build these systems to make profits [4]. A large portion of the private-for-profit sector developed its secret risk assessment algorithms for the public duties of the criminal justice system. This relationship, however, risks the criminal justice system's independence and transparency [58]. It is easier for the government to build people's trust than the private sector [11]. Therefore, the private sector is working to stand on the shoulders of the government to make up for the lack of trust people have in their activities. In addition, there are concerns about the unreliability of these risk assessment tools developed by private sectors [13]. These concerns are related to most risk assessment tools lacking inclusion in their development. Most private sectors exclude the insight, knowledge and close collaborative work of criminal justice systems officials when developing these systems. Hence, it is crucial to include criminal justice system officials in developing, implementing, and using risk assessment tools [6].

There is confusion about the inter-relatedness of explainability of risk assessment algorithms for recidivism prediction and its inter-relatedness with other ethical requirements such as accountability and transparency. A review of algorithmic explainability and legal reasoning [59] stated that the explainability of the recidivism risk assessment tools is obscure and needs further research. In addition, the literature emphasized that explainability should be treated as a formal-procedural criterion separated and distinguished from its closely related ethical requirements (accountability and transparency) and should be applied exclusively to machine outputs and decisions. Zsolt Zodi [59] argued that explaining algorithmic decisions through common sense is often tricky because of the massive gap in statistical explanations. Therefore, explainability should be applied at every algorithmic decision-making point and translated into human language for a better reason. But it is best to avoid using these recidivism risk assessment tools when deciding people's rights.

Accountability is rarely defined and addressed in the risk assessment literature for predicting recidivism risk. Accountability is, however, needed when things go wrong with assessment systems used for predicting the risk of recidivism [8]. There is an eminent call for recidivism risk assessment algorithmic design [58,60]. These algorithms are gaining ground in the evolution of the criminal justice system. Hence, who is accountable for the algorithm's predictive decisions when they go wrong? Is it the organization, programmers, or the stakeholders [61]?

Introducing an algorithmic decision-making system into the criminal justice system brought concerns about lack of accountability and equal protection [14,47]. In times past, some criminal justice managers have leveraged these predictive tools as vulnerable tools for blame shift when their decisions may harm a convicted individual or a social group.

Risk assessment tools' suitability for predicting recidivism risk needs more attention in different cases. This attention falls on critical ethical concerns escalating from deploying a large-scale of these systems [62]. The relationship between different ethical dimensions, actions of involved stakeholders to tackle ethical problems, and possible ways to improve the recidivism risk assessment system's development should be ethically resolved [15]. Furthermore, legislators must implement laws and regulations to curb the activities these algorithms can automate to attain accountability and achieve transparency [15,60]. In conclusion, developers should build recidivism risk assessment tools as human-in-the-loop systems for effective oversight, interactions, and trust in decision-making algorithms [63,64].

In this paper, we conducted a systematic literature review of research works tackling one or more of the requirements to achieve a trustworthy AI for predicting the risk of recidivism. This review revealed that no research focuses on AI's trustworthiness in predicting recidivism risk. To the best of our knowledge, our study is the first to consider the essential requirements for an AI to be considered trustworthy for predicting the risk of recidivism in the criminal justice system.

4. Methodology

4.1. Review Technique

Our methodical literature survey technique had three phases: (i) actively planning, (ii) conducting and reporting the review results, and (iii) exploration of research challenges. The systematic survey described in this paper followed the widely accepted guidelines and process outlined in [65,66]. The remainder of this section details the research questions, the process for identifying research, and the data extraction process.

4.2. A. Research Questions

The following are the identified research questions for this review:

- Q1: Are there works proposing the use of AI for predicting recidivism? How many?
- Q2: How many of these works considered the ethics and trustworthiness of the AI system?
- Q3: What are the essential requirements of trustworthy AI for predicting the risks of recidivism?
- Q4: What challenges hinder the development of trustworthy AI for predicting the risk of recidivism?

4.3. B. Search Strategy

The identification of research started by formulating a search query (Figure 1) based on the research focus, "Ethics and Trustworthiness of AI in Criminal Justice System when Predicting the Risk of Recidivism". The search query entails three parts: (1) All ML algorithms, including artificial neural networks algorithms; (2) terminologies used in the judiciary system; and (3) The requirements of trustworthy AI based on European Commission and Ethics guidelines for trustworthy AI. The Web of Science was chosen as the research database and was queried based on the formulated query and metadata (Author, Title, Source, and Abstract). Also, IEEE, Springer Nature, Assoc Computing Machinery, Elsevier, Mdpi, Wiley, Sage, and Taylor & Francis were the selected publishers. The research papers were from the year 2010 to mid-year 2023.

The search query returned 1826 research papers. In line with Kitchenham [66], two researchers met regularly to screen the 1826 papers based on titles and abstracts. All papers not addressing any of the requirements of trustworthy AI concerning predicting the risk of recidivism were removed. The two researchers removed duplicated papers and all papers unrelated to the research scope. This procedure reduced the number of papers from 1826 papers to 49 papers. In the next pilot phase, we started reading the 49 papers. We found other beneficial papers cited in the 49 papers during the reading. Therefore, we used a snowballing technique: looking at the references in all 49 papers and carefully selecting related research papers. The selected papers added up to 58 research papers from

the snowballing process. Next, we removed unrelated and duplicated papers based on reading the title and abstract of the 58 snowballed papers. The inclusion/exclusion criteria are shown in Table 1). Through this process, we obtained a total of 20 papers. Therefore, in addition to the initial 49 papers, we had 69 papers as the relevant papers for the systematic review of our research scope.

```
('Artificial Intelligence' OR 'AI' OR 'Machine Learning' OR 'ML' OR
\b'Supervised learning' OR 'Unsupervised learning' OR  'Reinforcement
Learning' OR 'Deep Learning' OR 'Neural networks' OR 'Automated Decision
Making' OR 'Fair ML' OR 'Algorithmic Fairness' OR 'Interpretable AI' OR
'Trustworthy AI' OR 'Algorithmic Decision Making' OR 'Robot Judge')
AND
('Crime' OR 'judgment' OR 'punishment' OR 'Criminal Recidivism' OR
'Criminal Justice' OR 'Criminal System' OR 'Judicial Intelligence' OR
'Parolee' OR 're-offence' OR 'Legal' OR 'Criminal Sentencing' OR 'Judicial
analytic' OR 'Recidivism')
AND
('trust' OR 'bias' OR 'ethics' OR 'fairness' OR 'interpretability'
OR 'transparency' OR 'diversity' OR 'explainability' OR 'robustness'
OR 'accountability' OR 'safety' OR 'privacy' OR 'Human oversight' OR
'Environmental well-being' OR 'stability' OR 'non-discrimination' OR
'Societal well-being')
```

Figure 1. Search String.

Table 1. Inclusion and exclusion criteria considered.

Inclusion Criteria	Exclusion Criteria
Full Text	Duplicated Studies
Articles written in English relating to the Trustworthiness of AI when predicting the risk of recidivism	Non-English
Published between 2010 to 2023	Published before the 2010
Published in Conferences, Journals, or Books	Uncompleted studies

4.4. Data Extraction

The data extracted from the selected 69 papers were bibliographic, the paper's objective, the methodology used, the dataset used, the result obtained, types of ML used, the dataset location, and the statistical measures understudied. We entered these data into a spreadsheet. However, we made known little of the extra data in this paper.

4.5. Quantitative Analysis

We carried out a quantitative analysis of the 69 papers. From our review, very few research works focus on the concept of trustworthy AI systems in predicting the risk of recidivism. Most of these studies focus on analyzing and surveying AI's bias and fairness, with few focusing on AI's accuracy, interpretability, and transparency in the criminal justice system when predicting offenders' recidivism risk. Others address the possible impacts of AI applications in predicting the risk of recidivism. All these studies are fragments of the whole concept of our research scope.

There is an increasing concern about, and need to evaluate, the requirements of trustworthy AI systems when utilized to assess recidivism risk in the criminal justice system. Our analysis shows significant research relating to the ethics and trustworthiness of AI when predicting the risk of recidivism began in 2016, as seen in Figure 2. Very few studies were carried out from the year 2010 to 2015. It is worth noting that, until mid-2023, seven papers related to trustworthy AI for predicting recidivism risk were published. In

conclusion, most literature evaluating AI's impact when predicting the risk of recidivism is US-based.

There are 50 publication venues found in our literature review. The two major publication venues were Artificial Intelligence and Law and the Journal of Quantitative Criminology.

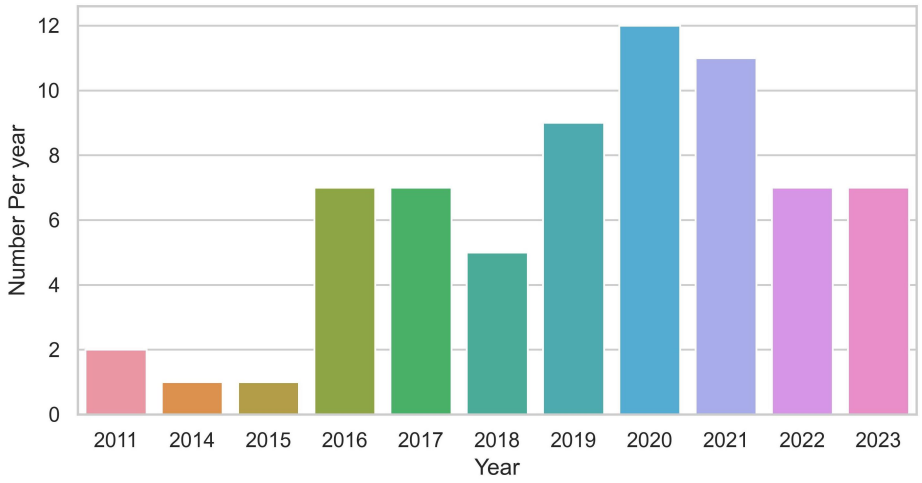

Figure 2. Number of papers per year.

Regarding publications, a significant proportion (45 out of 69) were journal articles, 23 were conference papers, and one was a book (see Figure 3). We classified each of the 69 papers as either technical or theoretical. In Section 3, we discuss 33 theoretical papers as related works. Section 6 discusses the 36 technical papers addressing the experimental approach to achieving the ethics and trustworthiness AI requirements when assessing recidivism risk as an extended analysis. The 36 papers were separated and discussed based on the requirements of trustworthy AI they are addressing in Section 6.

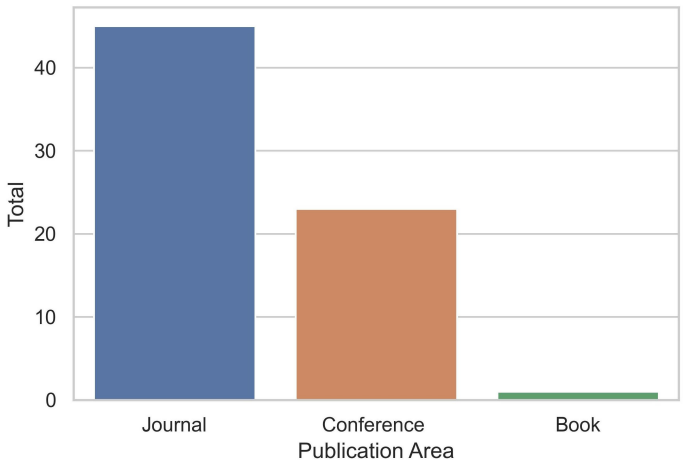

Figure 3. Type of publication.

5. Proposed Extended Requirements for Trustworthy AI in Criminal Judiciary System

Using the seven EC Requirements for Trustworthy AI as a baseline (see Section 2.3), we proposed an extension of the requirements that will facilitate the Ethics and Trustworthiness of AI when predicting the risk of recidivism in the criminal justice system (see Figure 4). We added four more requirements to the baseline following a thorough literature review. The proposed requirements are consistency, reliability, explainability, and interpretability. These four proposed requirements make the total requirements for trustworthy AI in predicting the risk of recidivism to be 11 requirements. This section gives an overview of these four added requirements.

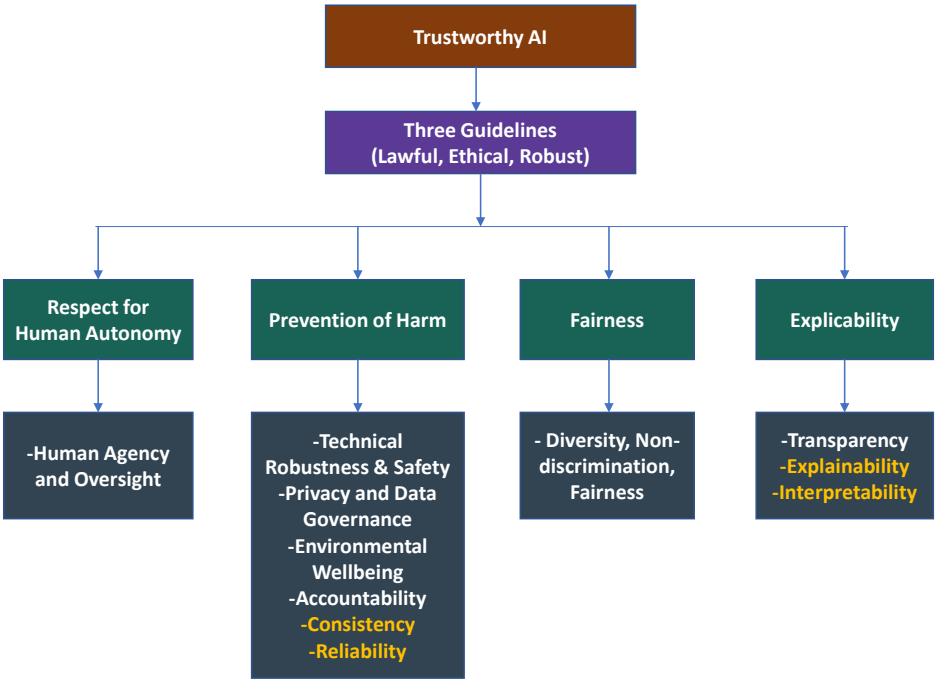

Figure 4. Extension of trustworthy AI framework.

5.1. Consistency

Consistency is a must-consider requirement in recidivism risk assessment tools. Recidivism risk assessment tools predict an offender's possibility of cycling back into the criminal justice system. Therefore, it will be a justifiable decision that criminals with alleged similar offences receive the same assessment score irrespective of demographic attributes or features [67]. In addition, removing or adding a feature in recidivism risk assessment tools should not influence the predictive score of offenders. Consistency ensures the predictive tools can maintain the same predictive score for each offender at all times, irrespective of changes in the offender attributes. A study on the auditing black-box models by obscuring features gave a straightforward process to the impact of features on predictive models' consistency [68]. Regarding decision makers using recidivism risk assessment tools as an assistant, how consistent are they in deciding offenders' risk of recycling back into the criminal justice system? The consistency requirement will help explore these areas more.

5.2. Reliability

The reliability of recidivism risk assessments is poorly understood. [9,69]. Reliability revolves around the internal design and structure of a predictive algorithm, the decision-makers involved in predicting the fate of offenders, and if the risk factors identified in a particular jurisdiction can be applied in other jurisdictions [69–71]. A study on the impact of ML risk forecasts of recidivism provided an understanding of reliability as an algorithm can consistently make corresponding forecasts for given offenders irrespective of random variations built into an algorithm [72]. This can be associated with the difference or change in the hyper-parameters used in training the algorithm. The uniformity of the risk assessment algorithm in forecasting an offender's recidivism risk will enhance the algorithm's reliability and ensure the user's trust [69].

5.3. Explainability

The explainability of predictive models focuses on the endpoint of the AI systems used to assess the outcome of an offender's recidivism. This concept has been greatly confused with AI models' transparency in predicting recidivism. The process or procedure of an AI system design may be transparent but challenging to explain. It is important to note that it is an obstacle for ML tools to explain their decisions as human beings would. Therefore, it is

imperative to include explainability as a stand-alone requirement of trustworthy AI when predicting recidivism. Explainability focuses on understanding why and what underlying factor constitutes the existence of the specific decision. Explainability is constructively examined in [4,59,73].

5.4. Interpretability

Interpretability is a dire concern in the criminal justice system when predicting the likelihood of recidivism [4,21,54,68,74–80]. Interpretability emphasizes the model computation being intuitive and meaningful to human understanding. Interpretability excludes the knowledge of data input or how the data relate to the outcome but how the model uses the different datasets to decide or make predictions about offenders and its relationship with ethical principles. Therefore, if the Interpretability of the AI models is unclear, it may be difficult for its users to trust in assessments of an offender's risk of recidivism.

6. Extended Analysis

This section enumerates and discusses the different practical approaches to achieving trustworthy AI systems that predict recidivism risk, summarized in Table 2. These discussions are based on the literature addressing the technical implementation of the requirements of trustworthy AI when predicting recidivism risk and the important findings we found.

AI in the criminal justice system for predicting recidivism is not new. However, there have been trust issues in applying these intelligent tools. This literature review revealed 11 requirements to achieve AI's ethics and trustworthiness for predicting recidivism risk. A point to note is that some of the practical papers address more than one requirement. Twenty-one out of thirty-six practical papers focused on the concept of fairness, which indicates that fairness is the biggest concern of criminal justice system stakeholders and scholars. From Figure 5, interpretability is another significant aspect in achieving trustworthy AI algorithms. However, these two can not stand alone to fully ascertain and draw people's trust in using risk assessment tools to predict recidivism risk. Less to no work has been undertaken regarding the other requirements. Therefore, there is a dire need to address the other requirements and make a balance across all the requirements. In the subsequent section, we briefly discuss the different requirements that have received less to no work in a subsection.

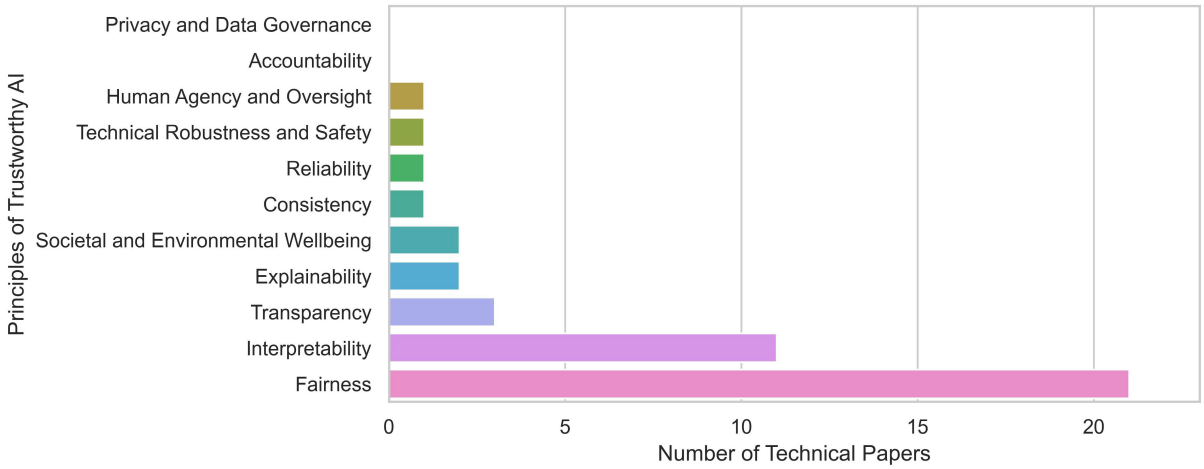

Figure 5. Requirements of trustworthy AI when predicting the risk of recidivism.

6.1. Fairness

Fairness is the most studied and focused-on requirement by criminal justice system stakeholders when applying the AI system to predicting recidivism risk. There have been concerns about the ability of AI models to bring equity, uniformity, and non-discriminating impacts to any individuals, groups, or races in a population. The primary cause of this concern is bias in AI models for predicting recidivism. There has been evidence of biases

in the data used in training AI models to predict the recidivism risk, which impacts the fairness of AI model outcomes [54,55,81]. The second concern is the trustworthiness of the individuals developing these AI models for predicting recidivism risk. How do we ascertain that these individuals are not building these models for selfish interests? The trustworthiness of developers is an important aspect that needs to be looked into to achieve trustworthy AI systems for predicting recidivism risk. Another concern is the inappropriate use of recidivism risk assessment models that do not conform to the domain in use and thereby cause unfairness in their predictions. The different criminal justice system stakeholders must consider these various concerns for fairness to be incorporated into AI models to predict recidivism risk.

On papers regarding the ethics and trustworthy AI in predicting recidivism risk, we came across 21 out of 36 papers working towards the practical approach of achieving fairness in AI models when predicting recidivism risk, as seen in Figure 5. From reviewing these papers, we discussed and enumerated our observations and findings that need consideration to achieve fairness when predicting recidivism risk. In addition, we carried out a comparative analysis on fairness as seen in Table 2.

A study on interpretable recidivism prediction [4] compared the fairness of different interpretable ML models with two currently used recidivism systems: COMPAS and Arnold Public Safety Assessment. The study built interpretable models to predict recidivism better than the two currently used recidivism systems. However, they face the challenge of their interpretable models needing to perform better when used on recidivism datasets from other jurisdictions or locations. The author points out that AI systems for predicting recidivism are often limited to the jurisdiction they are built for and cannot be used in other jurisdictions. In addition, a study on causes of algorithmic bias in juvenile criminal recidivism [73] pointed out that most of the research assessing the fairness of AI models on offenders assessed for recidivism risk is United States based. The study indicated a dire need to collect more new datasets from other locations, and researchers must focus more on different jurisdictions.

Furthermore, we observed that many researchers tend to apply fairness definitions generally defined in other domains to the criminal justice system when predicting the risk of recidivism. However, it is crucial to understand that many of the available fairness criteria are irrelevant to the criminal justice system [4]. A case study on predictive fairness to reduce misdemeanour [16] also affirms this claim that AI models for predicting recidivism risk do not satisfy all definitions of fairness. Therefore, there is a need to explore fairness definitions relevant to applying AI models to predicting recidivism risk for more equity, efficiency, and effectiveness of these AI models. Another issue related to fairness in AI models for predicting the risk of recidivism was identified in a study using bias parity score to find feature-rich models with the most negligible relative bias [51]. The identified issue is the problem of sensitive features (such as race, gender, mental health status, weight, etc.), which could impact the fairness of AI models on offenders. The author [51] emphasized that there is a need for further study of the effect of these protected features and the development of new metrics to evaluate the fairness of AI models. A survey on algorithms addressing trade-offs in predicting recidivism [82] points out that there is rare research on the impact of risk assessment models on racial disparities. In the study, the author claims that recidivism risk assessment models have led to the release of the low-risk offender but often do not consider racial equality. Therefore, racial equality is a fairness issue that should be addressed when predicting recidivism risk.

Regarding the fairness of recidivism risk assessment tools, prior probability shifts often occur in recidivism datasets and affect fairness. A prior probability shift is where the training and test set distributions for developing AI models differ. A study establishing fairness under prior probability shifts [83] proposed a Combinatorial Algorithm for Proportional Equality (CAPE) method and introduced a prevalence difference metric to solve the problem of prior probability shifts. The author claimed that their introduced metrics outperform all other available metrics. Therefore, it will be advisable to explore the CAPE method to validate its effects when predicting recidivism risk.

Table 2. Comparative analysis of fairness.

Ref	Purpose of Paper	Dataset	Algorithm Model Used or Studied	Evaluation Metrics	Location
[4]	Analyzed the performance of interpretable models regarding prediction ability, sparsity, and fairness	Broward County, Florida dataset & Kentucky dataset	LR, SVM, RF,DT, CART, Explainable Boost Machine	FPR, FNR, Accuracy	USA
[73]	Investigate the recidivism risk on general-purpose ML algorithms, at the expense of not satisfying relevant group fairness metrics	Juvenile Justice System of Catalonia	LR, Multi-Layer Perceptron, SVM with a Linear kernel, KNN, RF, DT, NB	Balanced accuracy, TPR, TNR, AUR-ROC	Spain
[81]	Studied and design Singular Race Models for recidivism and its effects on the accuracy and bias	Florida Department of Corrections (FDOC) & Florida Department of Law Enforcement	KNN, RF, AdaBoost, DT, SVM, ANNs	Accuracy, FPR, FNR	USA
[51]	Introduced a new fairness measure and an enhanced feature-rich representation that permits the selection of the lowest bias models	Recidivism of Prisoners Released in 1994	ANN	TP, TN, FP, FN, PPV	USA
[74]	Investigated the use of Mugshots to address racial disparity	Miami-Dade County Clerk of the Court	MTCNN	Accuracy	USA
[83]	Proposed a method called CAPE to solve fair classification problems in the presence of prior probability shifts.	Broward County, Florida (COMPAS) and MEPS	CAPE	Prevalence Difference (PD), Proportional Equality (PE)	USA
[82]	Compared three strategies for debiasing algorithms and how they affect the fairness trade-off when predicting recidivism.	Federal Probation System—Post Conviction Risk Assessment (2009–2019)	Post Conviction Risk Assessment (PCRA) algorithms	AUC, PPV, FPR, FNR	USA
[54]	Illustrated the construction of officer risk assessment modelling using the demographic, network, and Hawkes point process features.	Use of Force Complaint data from the Chicago Police Department	Boosted Decision Tree, Feed-Forward NN, Auto Machine Learning	AUC, Logloss, RMSE, MAE	USA
[84]	Introduced three fairness definitions that satisfy intersectional fairness, desiderata, differential fairness and its bias amplification	Broward County, FL (COMPAS), UCI Adult data repository (1994)	Neural Network (ADAM)	Accuracy, F1-Score, AUC-ROC	USA
[53]	Introduced a novel probabilistic formulation of data preprocessing for reducing discrimination	Broward County, Florida (COMPAS)	LR, RF	AUC	USA
[85]	Developed an ML model predicting the criminal offence type committed in a large transdiagnostic sample of psychiatry patients	Ontario Review Board (Forensic mental health system)	RF, SVM (Radial Kernel), XGBoost	Accuracy, AUC-ROC, Sensitivity, Specificity, Confidence Interval	Canada

Table 2. *Cont.*

Ref	Purpose of Paper	Dataset	Algorithm Model Used or Studied	Evaluation Metrics	Location
[23]	Applied fairness criterion originating in educational and psychological testing to assess the fairness of recidivism prediction instruments.	Broward County, Florida		FPR, FNR, PPV	USA
[19]	Analyzed ProPublica on the risk assessment tool (COMPAS)	Broward County, Florida (COMPAS)	LR	AUC-ROC, FN, FP, Sensitivity, Specificity, PPV, NPV	USA
[68]	Presented a technique for fairness and auditing black-box models using a variety of publicly available datasets and models	National Archive of Criminal Justice Data	SVM, Feedforward NN	Gradient Feature Auditing (GFA)	USA
[86]	Compared the fairness predictions of risk assessment tools and humans	Broward County, Florida (COMPAS)	LR, Nonlinear SVM, COMPAS software	Accuracy, FP, FN	USA
[87]	Proposed an approach to increase recidivism prediction accuracy while reducing race-based bias	Recidivism of Prisoners Released in 1994	XGBoost, LR, SVM	Accuracy, FPR parity, FPR, FNR, TPR, TNR, Monte Carlo cross-validation	USA
[88]	Research on human interactions with risk assessments through a controlled experimental study on Amazon Mechanical Turk	U.S. Department of Justice	Gradient Boosted Trees	AUC-ROC, Accuracy, FPR	USA
[24]	Studied and compared the accuracy and fairness of risk assessment tools and humans in predicting recidivism risk	Broward County, FL (COMPAS)	LDA, LR, Non-Linear SVM, COMPAS	AUC-ROC, Accuracy, FPR, FNR	USA
[16]	Addressed definitions and metrics for fairness that exists in the literature to optimizes public policy problem	City Attorney's case Management System	Regularised LR, Decision Trees, RF, Extra Tree Classifiers	FPR, FDR, FOR, FNR, FP, FN, Recall	USA
[89]	Designed a system that algorithmically redacts race-related information to reduce potential bias	American District Attorney's Office	Gradient-Boosted Decision Tree	Accuracy, AUC-ROC, FPR, FNR	USA
[90]	addressing the shortcomings of bias-error trade-off in AI algorithms	1994 Census Income dataset, German Credit Dataset	Adaptive Boosting, SVM, LR	FPR, FNR	USA

In conclusion, research on algorithmically masking race in charging decisions [89] pointed out the ethical implications of penalizing offenders for possible future misconduct that they have not committed. The author claimed that assessing recidivism risk would benefit society; however, criminal justice system stakeholders should conduct a more in-depth study on the forward-looking predictive models for predicting recidivism risk. Lastly, it is crucial to further study the issue of fairness both technically and in non-technical settings for AI models used for assessing the risk of recidivism to be widely accepted by all stakeholders [88].

6.2. Interpretability

Interpretability is the understanding of AI models for predicting recidivism risk, their computation, and how it relates to the model outcome. Interpretability as a requirement of a trustworthy AI framework when predicting recidivism risk surged when AI began to predict sensitive issues of humans' risk of recidivism. Scholars began challenging the computation process of these recidivism risk assessment models and are seeking answers on interpreting their workings and how they arrive at their decisions. Scholars and researchers have theoretically challenged using AI models to predict recidivism risk. However, a few technical approaches worked towards understanding the interpretability of AI models predicting recidivism risk (see Table 3).

One of the technical approaches carried out in [4] studied interpretable ML models for recidivism prediction by generating multiple interpretable models and black-box models, which the author compared based on prediction ability, fairness, and also against two state-of-the-art models (Arnold Public Safety Assessment, COMPAS). The author pointed out an ML method known as the Superspace Linear Integer models (SLIM) model as being an excellent interpretable model that does not violate fairness definitions when predicting recidivism [4,80]. However, the author also pointed out that the main drawback of the SLIM model is the difficulty in solving the integer programming problem. Therefore, it will be advisable to explore further the SLIM model on several datasets to validate its interpretability when predicting recidivism risk.

Research on detecting racial inequalities in the criminal justice system [74] proposed possible ways to address racial disparities when predicting recidivism and proposed a practical approach to understanding the interpretability of deep learning models for predicting recidivism risk. The author used mugshots to train its deep learning models to understand how the models recognize racial categories and how the model can fill in missing race data in the recidivism dataset. The experimental approach followed the design and methodology in standard facial processing technology pipelines to address different sources of deep learning model bias. The use of mugshots in categorizing race is a novel approach yet to be fully explored. This study [74] achieved a significant accuracy in categorizing defendants' race using mugshots and helping mitigate racial disparities. However, the author stated their method must be extended to other jurisdictions to examine its efficiency.

Among the 11 technical papers that touch on the concept of Interpretable models [4,21,54,68,74–80], the standard ML algorithm studied are logistic regression, classification and regression tree, support vector machine, neural networks and superspace linear integer models (SLIM), with the datasets most used for understanding the interpretability of these algorithms being the Broward County, Florida dataset, USA, and the Chinese AI and Law (CAIL) 2018 dataset, China. To the best of our knowledge, this suggests that more datasets must be collected to study the interpretability of AI models, especially in Europe.

In addition, a case study on criminal law based on multi-task learning [79] opined that for ease of interpretability of AI models for predicting recidivism, it is vital to include expert knowledge of relevant criminal justice system personnel when auditing and designing AI models to predict the recidivism risk. Lastly, a study on predicting domestic violence recidivism [78] pointed out two limitations in the literature. One, Sanuri et al. [78], pointed out the limited research on the interpretability of models when predicting domestic violence recidivism. Secondly, Sanuri et al. [78] pointed out that the ROC (receiver operating char-

acteristic) measure, which relies on detailed offender information, can help further in the interpretability of AI models used for predicting recidivism risk; however, there have been limitations to the use of ROC measures.

Table 3. Comparative analysis of interpretability.

Ref	Purpose of Paper	Dataset	Algorithm Model Used or Studied	Evaluation Metrics	Location
[4]	Analyzed the performance of interpretable models regarding prediction ability, sparsity and fairness	Broward County, Florida dataset & Kentucky dataset	LR, SVM, RF, DT, CART, Explainable Boost Machine	FPR, FNR, Accuracy	USA
[80]	Presented interpretable binary classification models to predict general recidivism as well as crime-specific recidivism	Recidivism of Prisoners Released in 1994	CART, LR, SVM, Stochastic Gradient Boosting (Adaboost)	TPR, FPR, AUC-ROC	USA
[21]	Achieved multi-granularity inference of legal charges by obtaining subjective and objective elements from the fact descriptions of legal cases	CAIL 2018	SVM, Deep Pyramid CNN, ELECTRA, QAjudge	Macro-Precision, Macro-Recall, Macro-F1	China
[74]	Investigate the Interpretable model through the use of mugshots to racial bias	Miami-Dade County Clerk of the Court	MTCNN	Accuracy	USA
[79]	Used Multi-task learning to conduct joint training with the task of crime prediction	CAIL 2018	LibSVM, LSTM, Multi-Label-KNN, BiLSTM	Precision, Recall, F1-measure, F-macro, F-Micro	China
[54]	Illustrated the construction of interpretable risk assessment modelling using demographic features	Use of Force Complaint data from the Chicago Police Department	Boosted DT, Feed-Forward NN	AUC, Logloss, RMSE, MAE	USA
[78]	Employ Decision Tree induction to obtain both interpretable trees as well as high prediction accuracy	NSW Bureau of Crime Statistics and Research (BOCSAR) Re-offending Database	DT, LR	AUC-ROC, TPR, FPR	Australia
[77]	To establish open-source algorithms as the standard in highly consequential contexts that affect people's lives for reasons of transparency and collaboration.	Broward County Florida	Ridge Regression, LASSO Regression, Elastic Net Regression	AUC-ROC,	USA
[76]	Presented a method (Gradient Feature Auditing) to evaluate the effect of features in a data set on the predictions of models	National Archive of Criminal Justice Data	Deep NN, SVM, DT, Superspace Linear Integer Models (SLIM)	Balanced Classification Rate (BCR)	USA
[68]	Presented a technique for auditing black-box models using a variety of publicly available datasets and models	National Archive of Criminal Justice Data	SVM, Feedforward NN	Gradient Feature Auditing (GFA)	USA
[75]	A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent re-offending	Prison Service Inmate Information System and Central System Database	LR, CART, Multi-Layer Perceptron NN	AUC-ROC, Accuracy	UK

6.3. Transparency

From our literature review, we discovered that there is often considerable confusion about what transparency and interpretability mean. At times, authors inter-change the use and concepts of interpretability for transparency. Therefore, it is necessary to distinguish between transparency and interpretability clearly.

Transparency emphasizes the development procedures of AI models for predicting recidivism risk and making these development procedures available to criminal justice system stakeholders. However, making these development procedures available must satisfy ethical principles. At the same time, interpretability focuses more on the internal computation of AI models and how they relate to the model outcomes. Transparency is about documenting, communicating, and making the procedures for building AI models available without infringing on fundamental human rights and related ethical principles detrimental to the stakeholders. Transparency will enable criminal justice system stakeholders to challenge AI models' scientific validity for predicting recidivism risk.

From Figure 5, only three research papers worked toward transparency, but only one focused on understanding transparency. A study on open-source development of predictive algorithms [77,80] suggested that for transparency incorporation in the predictive algorithms of recidivism used in the criminal justice system, it is essential to pursue open-source algorithm development. Open-source algorithm development will help improve

predictions and lower the cost of judicial decision-making on offenders when released on parole, probation or after jail time. In addition, it will create room for public trust and open opportunities for performance improvement of recidivism predictive models. However, some challenges must be addressed before incorporating open-source development on AI models to predict recidivism risk. These are privacy issues and a lack of continuous support for open-source systems. The dataset used in training AI models contains sensitive information about criminal records. Therefore, there is a dire need for ethical rules that needs to be in place to address the privacy of defendants. Lastly, it is a fact that closed-source systems often attract financial contributions from those needing access to the closed-source systems. However, this is not the case for most open-source systems. Therefore, there is a need to devise a strategy or mechanism to ensure the continued financial support of open-source systems when predicting recidivism risk. In summary, more considerable work need addressing the concept of transparency and the different approaches to incorporate transparency into the design of AI models for predicting recidivism risk.

6.4. Other Requirements for Trustworthy AI

This section discusses the other essential requirements of trustworthy AI in predicting the risk of recidivism, such as privacy and data governance, human agency and oversight, technical robustness and safety, accountability, reliability, consistency, explainability, and societal and environmental well-being. From Figure 5, it is evident that the different requirements discussed in this section need much research to address them.

Our review found no research study tackling privacy and data governance when assessing defendants for recidivism risk. Data are essential to developing recidivism risk assessment models; however, there is a dire need to devise mechanisms to handle and protect offenders' data appropriately. This is crucial to gain the trust of criminal justice systems stakeholders in recidivism risk assessment tools. In addition, ensuring offenders' privacy conforms to ethical principles is a must. These offenders deserve a right to fundamental human rights and data information protection. It is important to note that privacy and data protection are among the most vital requirements of trustworthy AI for predicting recidivism risk as it further extends into other trustworthy AI requirements such as transparency, accountability and safety. This extension is in the sense that when considering transparency in the design of recidivism risk assessment tools, how do we also ensure that we balance protecting offenders' information? Also, on account of a breach of privacy protection and offenders' data protection, who will be held accountable for such violations? These questions warrant prompt researchers and criminal justice system stakeholders' focus to ensure that the privacy and data protection of the offenders assessed for recidivism risk is guaranteed. In terms of safety, hackers can hack these risk assessment models. Hackers can access and modify offenders' data or manipulate risk assessment tools to give false outcomes that may lead to the release of high-risk offenders into the communities. Therefore, privacy and data governance are paramount in developing risk assessment tools for predicting recidivism risk in the criminal justice system. In addition, there is room for research in the safety of recidivism risk assessment models, as the literature on recidivism risk assessment tools' security is still qualitative [91]. It is imperative to guarantee that these algorithms are not susceptible to attacks of any form, such as cyber-attacks that could tamper with the privacy of criminals or the computational workings of the system. In situations of adverse attacks, there should always be a fallback plan [91].

Accountability is another requirement that has received little to no focus when assessing criminals for recidivism risk. Accountability focuses on who should be accountable for the decisions or circumstances surrounding the outcomes made by predictive recidivism models on offenders. There has been anxiety surrounding who takes responsibility if the decisions made by these predictive models go wrong. As much as we have asked who should be responsible for any wrong occurring in the outcomes of recidivism risk assessment tools, what about developers? Many developers are involved in developing recidivism risk assessment models and who should be accountable for a failure in the de-

velopment process. Accountability involves all stakeholders, including users, developers, and governing bodies. However, at what point should each of these stakeholders be held responsible for failure in the development procedure of recidivism risk assessment models or the decision made by the recidivism risk assessment models is a question that needs to be answered and further explored.

On the requirement of human agency and oversight, there have been ideas on the potential possible collaboration of AI models and humans, especially domain experts, when designing predictive recidivism models and these models deciding the recidivism risk of offenders. This is important to boost the confidence and reliance of individuals on risk assessment tools [64]. A research study on algorithm-in-the-loop fairness analysis in risk assessment tools [88] emphasized human-algorithm relationships to ameliorate human decisions rather than a total focus on how algorithms can better make their decisions. The study [88] noted the essential benefits that would spring up from the interactions between humans and AI models. The author confirmed in their research that even though AI models can perform very well independently when making a decision, human-algorithm interaction will produce better decisions on offenders' recidivism risk. However, such personnel interacting with the recidivism assessment models when deciding whether an offender will recidivate must be trained and provided with adequate guidelines to avoid disparate interactions. As much as there is a dire need for extensive research on human agency and oversight as a requirement of trustworthy AI when predicting recidivism, a few future directions will be to look at the following. Firstly, what mechanisms will ensure human-algorithmic interactions when predicting recidivism and at what stage? Secondly, how do recidivism risk assessment tools impact criminal justice system stakeholders' sense of accountability when deciding an offender's recidivism risk? Lastly, there is a need for proper experimental application in real-world scenarios to ascertain the visibility of human-algorithm interactions when predicting recidivism risk.

Reliability is another concept introduced in a study evaluating the impacts of predictions on assessments of recidivism risks [72]. Reliability is the ability of recidivism risk assessment models always to make the exact prediction for a given offender, irrespective of variations made in the build-up of the predictive model. From our systematic review, this is the only literature to have dealt with understanding the relatedness of reliability when making predictions. From their findings, there needs to be concrete evidence to suggest whether it affects public safety. In addition, it is a concept that is important to validate the efficacy of predictive models' reliability in assessing risks of recidivism.

Furthermore, consistency is another aspect to look at when it comes to the issue of the trustworthiness of AI. Consistency is closely related to the reliability of predictive recidivism models. The difference is that it focuses on the variations of features used in training the predictive model [68]. Irrespective of these variations, we expect predictive models to maintain the same result in their outcomes for a given offender.

In our review, we found the importance of explainability as a vital requirement for trustworthy AI for predicting recidivism in two technical research works [4,73]. In [73], the author introduced an explainability method to show the negative consequences of unfairness mitigation techniques. In [4], the author described explainability as an approach that can provide an understanding of black-box models. However, the author [4] strongly argued that the concept of explainability has not yet fully reached a dependable level to fully provide the insights needed for black-box computations as commonly used in other scientific domains. Therefore, explainability should be limited to understanding AI decisions rather than its computational procedures, thereby keeping stakeholders abreast of the AI decisions [4]. The concept of explainability has not yet reached a dependable level to fully provide the insights needed for black-box computations as commonly used in other scientific domains.

In conclusion, exploring the mentioned and discussed requirements is imperative to ease the deployment of AI models in predicting the risk of recidivism in the criminal justice system.

7. Issues and Challenges

It is imperative to know that, despite the advantages and benefits of AI in predicting recidivism risk, there are challenges impeding these recidivism risk assessment models from being considered trustworthy. From our systemic review, we have identified a few issues that need the utmost attention.

7.1. Datasets Used

There has been an outcry about how the predictive algorithms used in predicting recidivism risk are biased and unfair. This outcry is associated with the fact that most datasets used are historical datasets that are not often updated. Due to time's evolution, the dataset influences the recidivism risk assessment model's outcomes. In addition, most datasets collected as recidivism datasets are compromised because many personnel involved in the collection of datasets do not implement due data collection procedures. Lastly, most datasets available are peculiar to one geographical location, especially the USA, which hinders their applicability in other jurisdictions worldwide. Therefore, there is a need for new datasets to be collected in Europe and other jurisdictions with careful consideration to avoid bias and be purpose-specific and updated regularly.

7.2. Standardization

From our review, it is essential to note that there is a lack of an agreed definition of several requirements of trustworthy AI for predicting recidivism. The lack of standard definitions of these requirements led to trade-offs among some of the requirements of trustworthy AI for predicting recidivism risk. As long as there are no instituted and concise definitions of these requirements, there will always be the problem of lack of trust. Trust comes from having a set agreed plan and a generally accepted view of a concept. Therefore, there is an imperative need for an evaluation method that cuts across all the developing cycles of predictive models of recidivism in the criminal justice system.

7.3. Metrics

Different metrics are available for evaluating AI algorithm performances. However, a metrics framework is needed to cut across the different requirements of trustworthy AI for predicting recidivism risk and developing recidivism risk assessment models. The choice of metrics will always be application-dependent. Therefore, a set of defined metrics for each requirement of trustworthy AI when predicting recidivism is needed. The metrics framework will further help validate how much criminal justice system stakeholders can trust recidivism risk assessment models.

7.4. Propensity to Trust: Private Sector vs. Public Sector

Timothy et al. [11] stated that it is easier for the government to build people's trust than the private sector. Therefore, the private sector tries to stand on the shoulders of the government to make up for the lack of trust people have in their activities. However, this relationship between the private sector and the government (public sector), stated by Timothy et al. [11], can be justified to a certain extent as it does not work for every society. There are instances whereby citizens in certain societies have significantly less trust, support and respect for the current government but have a high trust, support and respect towards the state institutions; citizens do not trust the government. In addition, the private sector sometimes defends the state institutions and even the government to protect their reputation, regulations and norms. This inter-relatedness between the citizens, public sector, and private sector poses the problem of who to trust when developing risk assessment systems for predicting the risk of recidivism.

8. Conclusions and Future Works

This paper reviewed the existing literature on or relating to the "Ethics and Trustworthiness of AI in Predicting the Risk of Recidivism in the Criminal Justice System". From our

review, several works of literature on the research scope focus on one or some part of the requirements of achieving a trustworthy AI for predicting recidivism risk in recent years. AI has been of great use in the criminal justice system and has come to stay. However, the criticism of AI system applications for predicting recidivism risk has brought concerns about the best approach to achieving a trustworthy AI that will be accepted and trusted by the people and its community.

In this paper, we extended the proposed seven EC requirements of trustworthy AI to eleven requirements to achieve a robust trustworthy AI system for predicting recidivism risk in the criminal justice system. These extended requirements are consistency, reliability, explainability, and interpretability. After thoroughly reviewing the existing literature, a future line of work is to ethically and technically explore the different trustworthy AI requirements in predicting recidivism risk at different development cycles of risk assessment systems. Exploring these requirements technically, we conducted an in-depth technical analysis of the fairness of AI in predicting the risk of recidivism in [92]. Other future works are to explore other requirements ethically and technically.

In several nations and jurisdictions where a risk assessment system is utilized to forecast recidivism risk, the causes for doing so are strikingly similar. However, considering the complexity of regulations and norms in law, ethics, and morals (as well as stereotypes in culture and traditions), this can impede the achievement of a universal ethical and trustworthy AI for predicting the risk of recidivism. Therefore, developing a recidivism risk assessment system should follow each jurisdiction's ethical policies and laws. Nevertheless, the different risk assessment system developers should consider the reliability (as a requirement of trustworthy AI) of the risk assessment system as it involves applying the risk factors identified in one jurisdiction to another.

The most commonly used dataset in the literature when predicting recidivism is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) associated with Broward County, Florida, in the United States of America. In our review, one of the hindrances of predictive algorithms predicting recidivism is that most datasets are peculiar to a geographical location, thus limiting their use in other jurisdictions. For this reason, more dataset collation that will suit the application of the predictive tools in each location is essential.

Furthermore, a generalized framework that will cut across all the requirements of trustworthy AI for predicting recidivism, which will help to structure the design and development of the risk assessment tools for predicting the risk of recidivism, is crucially needed. Likewise, research needs to be carried out on the hybrid collaboration of AI and human beings, as this will increase the efficiency in carrying out predictive tasks of recidivism. In conclusion, researchers must address and challenge the highlighted issues both ethically and technically.

Author Contributions: Conceptualisation, M.M.F.; methodology; M.M.F., I.T., M.B., R.C. and T.S.; data collection, M.M.F., I.T. and M.B.; resources, R.C.; discussion; M.M.F., I.T. and M.B.; writing—original draft preparation; M.M.F.; writing—review and editing, M.M.F., I.T., T.S. and M.B. All authors have read and agreed to the published version of the manuscript

Funding: This work was supported in part by the Science Foundation Ireland grants 13/RC/2094_P2 (Lero) and 13/RC/2106_P2 (ADAPT) and is co-funded under the European Regional Development Fund (ERDF).

Data Availability Statement: The data used in this review paper are derived from publicly available sources, such as published research articles, books, and publicly accessible databases. All references and citations are provided in the reference list of this paper. No new data were generated or collected for this review.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
FPR	False Positive Rate
FNR	False Negative Rate
TNR	True Negative Rate
TPR	True Positive Rate
PPV	Positive Predictive Value
NPV	Negative Positive Value
AUR-ROC	Area under the ROC Curve
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
ANN	Artificial Neural Networks
NB	Naive Bayes
LR	Logistic Regression
SVM	Support Vector Machines
CART	Classification and Regression Trees
MTCNN	Multi-task Cascaded Convolutional Network
RF	Random Forests
CAPE	Combinatorial Algorithm for Proportional Equality
XGBoost	eXtreme Gradient Boosted
DT	Decision Trees
LDA	Linear Discriminant Analysis

References

1. Sushina, T.; Sobenin, A. Artificial Intelligence in the Criminal Justice System: Leading Trends and Possibilities. In Proceedings of the 6th International Conference on Social, Economic, and Academic Leadership (ICSEAL-6-2019), Prague, Czech Republic, 13–14 December 2019; pp. 432–437. [CrossRef]
2. Kovalchuk, O.; Karpinski, M.; Banakh, S.; Kasianchuk, M.; Shevchuk, R.; Zagorodna, N. Prediction Machine Learning Models on Propensity Convicts to Criminal Recidivism. *Information* **2023**, *14*, 161. [CrossRef]
3. Berk, R.; Bleich, J. Forecasts of violence to inform sentencing decisions. *J. Quant. Criminol.* **2014**, *30*, 79–96. [CrossRef]
4. Wang, C.; Han, B.; Patel, B.; Rudin, C. In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *J. Quant. Criminol.* **2023**, *39*, 519–581. [CrossRef]
5. Mohler, G.; Porter, M.D. A note on the multiplicative fairness score in the NIJ recidivism forecasting challenge. *Crime Sci.* **2021**, *10*, 17. [CrossRef]
6. Cadigan, T.P.; Lowenkamp, C.T. Implementing risk assessment in the federal pretrial services system. *Fed. Probat.* **2011**, *75*, 30.
7. Green, B. The false promise of risk assessments: Epistemic reform and the limits of fairness. In Proceedings of the FAT* '20: 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 594–606. [CrossRef]
8. Lo Piano, S. Ethical principles in machine learning and artificial intelligence: Cases from the field and possible ways forward. *Humanit. Soc. Sci. Commun.* **2020**, *7*, 9. [CrossRef]
9. Desmarais, S.L.; Johnson, K.L.; Singh, J.P. Performance of recidivism risk assessment instruments in US correctional settings. *Psychol. Serv.* **2016**, *13*, 206. [CrossRef]
10. Green, B. "Fair" risk assessments: A precarious approach for criminal justice reform. In Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning, New York, NY, USA, 23–24 February 2018; pp. 1–5.
11. O'Loughlin, T.; Bukowitz, R. A new approach toward social licensing of data analytics in the public sector. *Aust. J. Soc. Issues* **2021**, *56*, 198–212. [CrossRef]
12. Bickley, S.J.; Torgler, B. Cognitive architectures for artificial intelligence ethics. *AI Soc.* **2023**, *38*, 501–519. [CrossRef]
13. Chugh, N. Risk assessment tools on trial: Lessons learned for "Ethical AI" in the criminal justice system. In Proceedings of the 2021 IEEE International Symposium on Technology and Society (ISTAS), Waterloo, ON, Canada, 28–31 October 2021; pp. 1–5. [CrossRef]
14. Hartmann, K.; Wenzelburger, G. Uncertainty, risk and the use of algorithms in policy decisions: A case study on criminal justice in the USA. *Policy Sci.* **2021**, *54*, 269–287. [CrossRef]
15. Alikhademi, K.; Drobina, E.; Prioleau, D.; Richardson, B.; Purves, D.; Gilbert, J.E. A review of predictive policing from the perspective of fairness. *Artif. Intell. Law* **2021**, *7*, 1–17. [CrossRef]

16. Rodolfa, K.T.; Salomon, E.; Haynes, L.; Mendieta, I.H.; Larson, J.; Ghani, R. Case study: Predictive fairness to reduce misdemeanor recidivism through social service interventions. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 142–153. [CrossRef]
17. Hamilton, M. The sexist algorithm. *Behav. Sci. Law* **2019**, *37*, 145–157. [CrossRef] [PubMed]
18. Dieterich, W.; Mendoza, C.; Brennan, T. *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*; Northpointe Inc.: Traverse City, MI, USA, 2016.
19. Flores, A.W.; Bechtel, K.; Lowenkamp, C.T. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probat.* **2016**, *80*, 38.
20. Hurlburt, G. How much to trust artificial intelligence? *It Prof.* **2017**, *19*, 7–11. [CrossRef]
21. Li, L.; Zhao, L.; Nai, P.; Tao, X. Charge prediction modeling with interpretation enhancement driven by double-layer criminal system. *World Wide Web* **2022**, *25*, 381–400. [CrossRef]
22. Zhang, Y.; Zhou, F.; Li, Z.; Wang, Y.; Chen, F. Fair Representation Learning with Unreliable Labels. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Valencia, Spain, 25–27 April 2023; pp. 4655–4667.
23. Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* **2017**, *5*, 153–163. [CrossRef]
24. Dressel, J.J. Accuracy and Racial Biases of Recidivism Prediction Instruments. Bachelor's Thesis, Dartmouth College, Hanover, NH, USA, 2017.
25. Kaur, D.; Uslu, S.; Rittichier, K.J.; Durresi, A. Trustworthy artificial intelligence: A review. *ACM Comput. Surv. (CSUR)* **2022**, *55*, 1–38. [CrossRef]
26. Emaminejad, N.; Akhavian, R. Trustworthy AI and robotics: Implications for the AEC industry. *Autom. Constr.* **2022**, *139*, 104298. [CrossRef]
27. Ma, J.; Schneider, L.; Lapuschkin, S.; Achtibat, R.; Duchrau, M.; Krois, J.; Schwendicke, F.; Samek, W. Towards Trustworthy AI in Dentistry. *J. Dent. Res.* **2022**, *101*, 1263–1268. [CrossRef]
28. Markus, A.F.; Kors, J.A.; Rijnbeek, P.R. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J. Biomed. Inform.* **2021**, *113*, 103655. [CrossRef]
29. Mora-Cantallops, M.; Sánchez-Alonso, S.; García-Barriocanal, E.; Sicilia, M.A. Traceability for trustworthy ai: A review of models and tools. *Big Data Cogn. Comput.* **2021**, *5*, 20. [CrossRef]
30. Kaur, D.; Uslu, S.; Durresi, A. Requirements for trustworthy artificial intelligence—A review. In *Advances in Networked-Based Information Systems*; Barolli, L., Li, K., Enokido, T., Takizawa, M., Eds.; NBiS 2020; Advances in Intelligent Systems and Computing; Springer: Cham, Switzerland, 2020; Volume 1264. [CrossRef]
31. Vining, R.; McDonald, N.; McKenna, L.; Ward, M.E.; Doyle, B.; Liang, J.; Hernandez, J.; Guilfoyle, J.; Shuhaiber, A.; Geary, U.; et al. Developing a framework for trustworthy AI-supported knowledge management in the governance of risk and change. *Lect. Notes Comput. Sci.* **2022**, *13516*, 318–333. [CrossRef]
32. Toreini, E.; Aitken, M.; Coopamootoo, K.; Elliott, K.; Zelaya, C.G.; Van Moorsel, A. The relationship between trust in AI and trustworthy machine learning technologies. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 272–283. [CrossRef]
33. Vincent-Lancrin, S.; van der Vlies, R. Trustworthy artificial intelligence (AI) in education: Promises and challenges. In *OECD Education Working Papers*; OECD Publishing: Paris, France, 2020. [CrossRef]
34. Ryan, M. In AI we trust: Ethics, artificial intelligence, and reliability. *Sci. Eng. Ethics* **2020**, *26*, 2749–2767. [CrossRef]
35. Connolly, R. Trust in commercial and personal transactions in the digital age. In *The Oxford Handbook of Internet Studies*; Oxford University Press: Oxford, UK, 2013; pp. 262–282. [CrossRef]
36. Beshi, T.D.; Kaur, R. Public trust in local government: Explaining the role of good governance practices. *Public Organ. Rev.* **2020**, *20*, 337–350. [CrossRef]
37. Smit, D.; Eybers, S.; Smith, J. A Data Analytics Organisation's Perspective on Trust and AI Adoption. In Proceedings of the Southern African Conference for Artificial Intelligence Research, Virtual, 6–10 December 2021; Springer: Berlin/Heidelberg, Germany, 2021; Volume 1551, pp. 47–60. [CrossRef]
38. Rendtorff, J.D. The significance of trust for organizational accountability: The legacy of Karl Polanyi. In Proceedings of the 3rd Emes-Polanyi Selected Conference Papers, Roskilde, Denmark, 16–17 April 2018; Roskilde University: Roskilde, Denmark, 2018.
39. Thiebes, S.; Lins, S.; Sunyaev, A. Trustworthy artificial intelligence. *Electron. Mark.* **2021**, *31*, 447–464. [CrossRef]
40. Liu, K.; Tao, D. The roles of trust, personalization, loss of privacy, and anthropomorphism in public acceptance of smart healthcare services. *Comput. Hum. Behav.* **2022**, *127*, 107026. [CrossRef]
41. Sutrop, M. Should we trust artificial intelligence? *Trames A J. Humanit. Soc. Sci.* **2019**, *23*, 499–522. [CrossRef]
42. High-Level Expert Group on Artificial Intelligence. In *Ethics Guidelines for Trustworthy AI*; European Commission: Brussels, Belgium, 2019. Available online: https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai (accessed on 3 July 2023).
43. OECD. *Tools for Trustworthy AI: A Framework to Compare Implementation Tools for Trustworthy AI Systems*; OECD Digital Economy Papers, No. 312; OECD Publishing: Paris, France, 2021. [CrossRef]
44. Floridi, L. Establishing the rules for building trustworthy AI. *Nat. Mach. Intell.* **2019**, *1*, 261–262. [CrossRef]

45. Janssen, M.; Brous, P.; Estevez, E.; Barbosa, L.S.; Janowski, T. Data governance: Organizing data for trustworthy Artificial Intelligence. *Gov. Inf. Q.* **2020**, *37*, 101493. [CrossRef]
46. Giovanola, B.; Tiribelli, S. Beyond bias and discrimination: Redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. *AI Soc.* **2023**, *38*, 549–563. [CrossRef]
47. Eckhouse, L.; Lum, K.; Conti-Cook, C.; Ciccolini, J. Layers of bias: A unified approach for understanding problems with risk assessment. *Crim. Justice Behav.* **2019**, *46*, 185–209. [CrossRef]
48. *ISO/IEC TR 24027:2021(E)*; Information Technology–Artificial Intelligence (AI)—Bias in AI Systems and AI Aided Decision Making. International Organization for Standardization, Vernier: Geneva, Switzerland, 2021.
49. Ireland, L. Who errs? Algorithm aversion, the source of judicial error, and public support for self-help behaviors. *J. Crime Justice* **2020**, *43*, 174–192. [CrossRef]
50. Berk, R. Accuracy and fairness for juvenile justice risk assessments. *J. Empir. Leg. Stud.* **2019**, *16*, 175–194. [CrossRef]
51. Jain, B.; Huber, M.; Elmasri, R.; Fegaras, L. Using bias parity score to find feature-rich models with least relative bias. *Technologies* **2020**, *8*, 68. [CrossRef]
52. Oatley, G.C. Themes in data mining, big data, and crime analytics. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2022**, *12*, e1432. [CrossRef]
53. du Pin Calmon, F.; Wei, D.; Vinzamuri, B.; Ramamurthy, K.N.; Varshney, K.R. Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis. *IEEE J. Sel. Top. Signal Process.* **2018**, *12*, 1106–1119. [CrossRef]
54. Khorshidi, S.; Carter, J.G.; Mohler, G. Repurposing recidivism models for forecasting police officer use of force. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Busan, Republic of Korea, 19–22 February 2020; pp. 3199–3203. [CrossRef]
55. Petersen, E.; Ganz, M.; Holm, S.H.; Feragen, A. On (assessing) the fairness of risk score models. *arXiv* **2023**, arXiv:2302.08851.
56. Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociol. Methods Res.* **2021**, *50*, 3–44. [CrossRef]
57. Grgic-Hlaca, N.; Redmiles, E.M.; Gummadi, K.P.; Weller, A. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In Proceedings of the 2018 World Wide Web Conference, Lyon, France, 23–27 April 2018; pp. 903–912. [CrossRef]
58. McKay, C. Predicting risk in criminal procedure: Actuarial tools, algorithms, AI and judicial decision-making. *Curr. Issues Crim. Justice* **2020**, *32*, 22–39. [CrossRef]
59. Ződi, Z. Algorithmic explainability and legal reasoning. *Theory Pract. Legis.* **2022**, *10*, 67–92. [CrossRef]
60. Mökander, J.; Juneja, P.; Watson, D.S.; Floridi, L. The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: What can they learn from each other? *Minds Mach.* **2022**, *32*, 751–758. [CrossRef]
61. Figueroa-Armijos, M.; Clark, B.B.; da Motta Veiga, S.P. Ethical perceptions of AI in hiring and organizational trust: The role of performance expectancy and social influence. *J. Bus. Ethics* **2022**, *186*, 179–197. [CrossRef]
62. Anshari, M.; Hamdan, M.; Ahmad, N.; Ali, E.; Haidi, H. COVID-19, artificial intelligence, ethical challenges and policy implications. *AI Soc.* **2023**, *38*, 707–720. [CrossRef] [PubMed]
63. Falco, G. Participatory AI: Reducing AI Bias and Developing Socially Responsible AI in Smart Cities. In Proceedings of the 2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), New York, NY, USA, 1–3 August 2019; pp. 154–158. [CrossRef]
64. Chiang, C.W.; Lu, Z.; Li, Z.; Yin, M. Are Two Heads Better Than One in AI-Assisted Decision Making? Comparing the Behavior and Performance of Groups and Individuals in Human-AI Collaborative Recidivism Risk Assessment. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; pp. 1–18.
65. Pai, M.; McCulloch, M.; Colford, J. *Systematic Review: A Road Map*, version 2.2; Systematic Reviews Group, UC Berkeley: Berkeley, CA, USA, 2004.
66. Kitchenham, B. *Procedures for Performing Systematic Reviews*; Keele Universit: Keele, UK, 2004; Volume 33, pp. 1–26
67. Ritter, N. Predicting recidivism risk: New tool in Philadelphia shows great promise. *Natl. Inst. Justice J.* **2013**, *271*, 4–13.
68. Adler, P.; Falk, C.; Friedler, S.A.; Nix, T.; Rybeck, G.; Scheidegger, C.; Smith, B.; Venkatasubramanian, S. Auditing black-box models for indirect influence. *Knowl. Inf. Syst.* **2018**, *54*, 95–122. [CrossRef]
69. Harada, T.; Nomura, K.; Shimada, H.; Kawakami, N. Development of a risk assessment tool for Japanese sex offenders: The Japanese Static-99. In *Neuropsychopharmacology Reports*; John Wiley & Sons: Victoria, Australia, 2023. [CrossRef]
70. Miller, C.S.; Kimonis, E.R.; Otto, R.K.; Kline, S.M.; Wasserman, A.L. Reliability of risk assessment measures used in sexually violent predator proceedings. *Psychol. Assess.* **2012**, *24*, 944. [CrossRef] [PubMed]
71. McPhee, J.; Heilbrun, K.; Cubbon, D.N.; Soler, M.; Goldstein, N.E. What's risk got to do with it: Judges' and probation officers' understanding and use of juvenile risk assessments in making residential placement decisions. *Law Hum. Behav.* **2023**, *47*, 320. [CrossRef]
72. Berk, R. An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *J. Exp. Criminol.* **2017**, *13*, 193–216. [CrossRef]
73. Miron, M.; Tolan, S.; Gómez, E.; Castillo, C. Evaluating causes of algorithmic bias in juvenile criminal recidivism. *Artif. Intell. Law* **2021**, *29*, 111–147. [CrossRef]

74. Dass, R.K.; Petersen, N.; Omori, M.; Lave, T.R.; Visser, U. Detecting racial inequalities in criminal justice: Towards an equitable deep learning approach for generating and interpreting racial categories using mugshots. *AI Soc.* **2022**, *38*, 897–918. [CrossRef]
75. Liu, Y.Y.; Yang, M.; Ramsay, M.; Li, X.S.; Coid, J.W. A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent re-offending. *J. Quant. Criminol.* **2011**, *27*, 547–573. [CrossRef]
76. Smith, B. Auditing Deep Neural Networks to Understand Recidivism Predictions. Ph.D. Thesis, Haverford College, Haverford, PA, USA, 2016.
77. Waggoner, P.D.; Macmillen, A. Pursuing open-source development of predictive algorithms: The case of criminal sentencing algorithms. *J. Comput. Soc. Sci.* **2022**, *5*, 89–109. [CrossRef]
78. Wijenayake, S.; Graham, T.; Christen, P. A decision tree approach to predicting recidivism in domestic violence. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Melbourne, VIC, Australia, 3–6 June 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–15. [CrossRef]
79. Yuan, D. Case Study of Criminal Law Based on Multi-task Learning. In Proceedings of the 2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE), Beijing, China, 23–25 October 2020; pp. 98–103. [CrossRef]
80. Zeng, J.; Ustun, B.; Rudin, C. Interpretable classification models for recidivism prediction. *J. R. Stat. Soc. Ser. A Stat. Soc.* **2017**, *180*, 689–722. [CrossRef]
81. Jain, B.; Huber, M.; Fegaras, L.; Elmasri, R.A. Singular race models: Addressing bias and accuracy in predicting prisoner recidivism. In Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments, Rhodes, Greece, 5–7 June 2019; pp. 599–607. [CrossRef]
82. Skeem, J.; Lowenkamp, C. Using algorithms to address trade-offs inherent in predicting recidivism. *Behav. Sci. Law* **2020**, *38*, 259–278. [CrossRef] [PubMed]
83. Biswas, A.; Mukherjee, S. Ensuring fairness under prior probability shifts. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Virtual, 19–21 May 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 414–424. [CrossRef]
84. Foulds, J.R.; Islam, R.; Keya, K.N.; Pan, S. An intersectional definition of fairness. In Proceedings of the 2020 IEEE 36th International Conference on Data Engineering (ICDE), Dallas, TX, USA, 20–24 April 2020; pp. 1918–1921. [CrossRef]
85. Watts, D.; Moulden, H.; Mamak, M.; Upfold, C.; Chaimowitz, G.; Kapczinski, F. Predicting offences among individuals with psychiatric disorders-A machine learning approach. *J. Psychiatr. Res.* **2021**, *138*, 146–154. [CrossRef] [PubMed]
86. Dressel, J.; Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.* **2018**, *4*, eaao5580. [CrossRef] [PubMed]
87. Jain, B.; Huber, M.; Elmasri, R.A.; Fegaras, L. Reducing race-based bias and increasing recidivism prediction accuracy by using past criminal history details. In Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments, Corfu, Greece, 30 June 30–3 July 2020; pp. 1–8. [CrossRef]
88. Green, B.; Chen, Y. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 90–99. [CrossRef]
89. Chohlas-Wood, A.; Nudell, J.; Yao, K.; Lin, Z.; Nyarko, J.; Goel, S. Blind justice: Algorithmically masking race in charging decisions. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, 7–8 February 2020; Association for Computing Machinery: New York, NY, USA, 2021; pp. 35–45. [CrossRef]
90. Fish, B.; Kun, J.; Lelkes, Á.D. A confidence-based approach for balancing fairness and accuracy. In Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, FL, USA, 5–7 May 2016; pp. 144–152. [CrossRef]
91. Zhang, S.; Yan, G.; Li, Y.; Liu, J. Evaluation of judicial imprisonment term prediction model based on text mutation. In Proceedings of the 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C), Sofia, Bulgaria, 22–26 July 2019; pp. 62–65. [CrossRef]
92. Michael, M.; Farayola, I.; Tal, S.T.; Connolly, R.; Bendechache, M. Fairness of AI in Predicting the Risk of Recidivism: Review and Phase Mapping of AI Fairness Techniques. In Proceedings of the 18th International Conference on Availability, Reliability and Security (ARES 2023), Benevento, Italy, 29 August–1 September 2023. [CrossRef]