Jinad *et al. J Surveill Secur Saf* 2024;5:19-35
DOI: 10.20517/jsss.2023.41

Journal of Surveillance, Security and Safety

Original Article

Open Access

Bias and fairness in software and automation tools in digital forensics

Razaq Jinad, Khushi Gupta, Ecem Simsek, Bing Zhou

Department of Computer Science, Sam Houston State University, Huntsville, AL 77340, USA.

Correspondence to: Dr. Bing Zhou, Department of Computer Science, Sam Houston State University, 1905 University Ave., Huntsville, AL 77340, USA. E-mail: bxz003@shsu.edu

How to cite this article: Jinad R, Gupta K, Simsek E, Zhou B. Bias and fairness in software and automation tools in digital forensics. *J Surveill Secur Saf* 2024;5:19-35. http://dx.doi.org/10.20517/jsss.2023.41

Received: 13 Nov 2023 **First Decision:** 22 Dec 2023 **Revised:** 4 Jan 2024 **Accepted:** 12 Jan 2024 **Published:** 26 Jan 2024

Academic Editor: Leandros Maglaras **Copy Editor:** Yanbin Bai **Production Editor:** Yanbin Bai

Abstract

The proliferation of software tools and automated techniques in digital forensics has brought about some controversies regarding bias and fairness. Different biases exist and have been proven in some civil and criminal cases. In our research, we analyze and discuss these biases present in software tools and automation systems used by law enforcement organizations and in court proceedings. Furthermore, we present real-life cases and scenarios where some of these biases have determined or influenced these cases. We were also able to provide recommendations for reducing bias in software tools, which we hope will be the foundation for a framework that reduces or eliminates bias from software tools used in digital forensics. In conclusion, we anticipate that this research can help increase validation in digital forensics software tools and ensure users' trust in the tools and automation techniques.

Keywords: Bias, fairness, digital forensics, automation, digital forensic software, digital forensics investigation

1. INTRODUCTION

In a recently released report by the National Police Chiefs' Council (NCPP), a United Kingdom's law enforcement body, 90% of criminal investigations now involve a digital element. Digital forensics is a crucial aspect of modern law enforcement, providing investigators with the ability to gather and analyze digital evidence in criminal investigations. As digital forensics develops, standardization and automation will become critical [1].

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

The use of software and automation tools in digital forensics has become increasingly prevalent in recent years, providing investigators with the ability to efficiently and effectively process large volumes of digital evidence. However, there is growing concern that these tools may introduce bias and unfairness into the forensic process. This is particularly concerning given the potential impact of forensic evidence on legal proceedings, where inaccurate or biased evidence can lead to wrongful convictions or acquittals.

Bias can simply be defined as the systematic and unjust treatment of individuals or groups. The basis for bias often lies in specific identifiable characteristics such as race, gender, or ethnicity. In the field of digital forensics, bias may arise at different points in the forensic process, encompassing stages such as data collection, analysis, and interpretation. For example, if a digital forensics tool is designed with algorithms that favor certain types of data or are not designed to detect certain types of evidence, this can result in biased outcomes. Acknowledging and addressing bias at each stage is crucial in ensuring the integrity and objectivity of the digital forensics process. This recognition helps forensic experts adopt methods that minimize bias, thereby enhancing the reliability of their findings and conclusions in the pursuit of accurate and fair investigations.

However, as the use of software and automation tools in digital forensics has increased, concerns have been raised regarding the potential for these tools to introduce bias and unfairness into the investigative process. To address these concerns, efforts are underway to develop standards and best practices for the design and use of digital forensics tools. These efforts include the development of ethical guidelines for digital forensics professionals, the establishment of standards for tool design and validation, and the promotion of transparency in tool design and use.

Bias and fairness in software and automation tools in digital forensics is a complex and important issue that requires careful attention. By addressing the sources of bias and implementing standards and best practices for tool design and use, we can work towards promoting fairness and accuracy in digital forensics and upholding the integrity of the legal system. The goal of this paper is to:

- Identify and categorize different manifestations of bias in digital forensics software tools and algorithms using real-life incidents.
- Explore several factors, such as technological constraints, programming errors, and inadequate or imprecise data, contributing to bias.
- Provide actionable solutions to address the identified challenges associated with bias in software automation in digital forensics.

The rest of this paper follows the following structure. In Section II, we start with some background information and discuss the existing work in the literature. Section III describes bias and its types. Section IV discusses digital forensics analysis using software and automation tools, while Section V lists the different real-life case scenarios of bias in digital forensics investigations. Furthermore, Section VI outlines our recommendations for mitigating bias. Finally, we discuss and provide our concluding remarks in Section VII.

2. BACKGROUND AND EXISTING LITERATURE

Digital forensics, traditionally defined as encompassing the collection, analysis, interpretation, and reporting of digital evidence, is a specialized field dedicated to investigating and analyzing digital evidence to uncover information in investigations. It encapsulates a systematic and thorough approach that includes collection, analysis, interpretation, and reporting of digital evidence[2].

The process of gathering digital evidence is a foundational step in digital forensics. This involves identifying, preserving, and securing digital information from various digital media. Analysis in digital forensics involves a thorough examination of collected digital evidence. Forensic experts use specialized tools and techniques to

scrutinize data for signs of criminal activity or other relevant information. The interpretation phase involves making sense of the analyzed digital evidence. Digital forensics experts interpret the findings in the context of the investigation, connecting the dots and drawing conclusions. This step requires expertise in understanding the significance of various digital artifacts and their implications for the case. Reporting is the final step in the traditional digital forensics process. Forensic professionals document their findings in a comprehensive report, detailing the methods used, the evidence analyzed, and the conclusions drawn.

To accomplish this, forensic software is employed to examine digital data that predominantly exists in an abstract form, with its electronic signal information being stored and interpreted via processing and translation. Although it is possible to manually examine data, this approach is not practical in most cases. Hence, the reliability and presentation of digital data, as facilitated by the digital forensics software tools employed in an investigation, are paramount for practitioners[3]. Undoubtedly, the significance of software tools in the realm of digital forensics cannot be emphasized enough.

Most of the phases in the digital forensics examination of evidence are completed by the software. The practitioner analyzes the results of the interpretation and examination of data from digital devices. However, the interpretation stages of digital storage media analysis cannot be manually verified as the content cannot be analyzed by sight alone. The problem arises when there is an incorrect interpretation of the data by the software. Without the capacity for visual or manual confirmation of the results, practitioners may be susceptible to misinterpretations or biases in forensic tools, possibly going unnoticed.

While tools are integral to ensuring precision across forensic disciplines, the reliance on software in digital forensics is arguably more substantial. In fact, practitioners would be incapable of accessing, acquiring, interpreting, and presenting digital data without the utilization of digital forensics software. Thus, the accuracy and reliability of the results generated by these tools are crucial in establishing the truth and supporting criminal justice proceedings. Any errors or biases during the examination process can jeopardize the entire investigation and compromise the evidential value of the results.

While experts in the digital forensics field acknowledge their dependence on digital forensics software, there is minimal discussion regarding bias within these tools and methods for its identification. The complicated nature and multitude of processes and procedures that forensic tools aim to handle render errors unavoidable. All software has "bugs" (minor programming anomalies) that can lead to inaccurate reports, presenting what seems to be factual information.

Additionally, within the legal domain, fairness is measured by ethical principles, the right to information, and the ability to challenge decisions[4]. Achieving algorithmic fairness necessitates a thorough comprehension of the interplay among potential factors that could have influenced a decision, encompassing counterfactual elements. However, investigators may ignore facts that contradict their perception, leading to erroneous inferences. If algorithmic analysis produces incorrect conclusions, it can undermine confidence in machine-generated results, which ought to be prevented.

In a study conducted by Sunde *et al.*[5], 53 digital forensics examiners analyzed an identical evidence file to explore the reliability and susceptibility to bias in their decision-making processes. The objective was to ascertain whether contextual information influenced the examiners' perceptions, interpretations, or findings and whether they consistently arrived at the same judgment when examining the identical evidence file with the same contextual details. The findings revealed that contextual information indeed introduced bias into the examiners' observations. Additionally, the study uncovered a limited level of consistency among examiners in their observations, interpretations, and conclusions. This diminished consistency underscores the necessity for quality assurance in digital forensics examinations to prevent incorrect findings from affecting the

investigation process.

Moreover, Stoykova[6] examines three classes of threats to fairness and the presumption of innocence that remain unaddressed in investigations. The first category involves the improper and inconsistent application of technology. The second category pertains to outdated procedural safeguards that are not tailored to contemporary processes and services in digital evidence. The third category concerns the absence of reliability testing in the practice of digital forensics. The article asserts that the presumption of innocence plays a vital role in bolstering evidence during the investigation phase, particularly when dealing with a substantial volume of heavily processed and corroborated data. Any utilization of technology in the investigation must conform to a reliability standard, carefully considering the potential for wrongful conviction and error. In conclusion, the article argues for legislative intervention, advocating for the enforcement of standards and validation procedures in the realm of digital evidence.

In[7], the authors aim to evaluate the current status of validation practices within digital forensics and to emphasize the requirement for further research in the area of discipline-specific method and tool validation. The authors suggest improving the assessment of tool limitations and constraining uncertainty associated with algorithms and their implementations. It is also proposed that validation efforts must aim to encompass all conditions known to cause errors.

Horsman[8] explores the present condition of digital forensic tool testing and the challenges associated with adequately evaluating applications for deployment in this field. The paper presents the fact that the field of digital forensics has not yet achieved a sufficient level of tool testing, and alarmingly, there are no apparent or effortless solutions at hand to address this situation. This concern is also depicted in the findings of the practitioner survey introduced in the research, and thus, the authors present possible solutions. Despite efforts made by organizations such as NIST, there has yet to be a widely embraced approach to tool testing in the field. The authors advocate for increased resources and standards to encourage greater participation and input from practitioners. This includes the development of a blueprint that defines the framework for dataset creation and subsequent digital forensics tool testing. Such a blueprint could standardize tool testing in digital forensics, fostering confidence in the outcomes produced.

Lastly, the study conducted in[9] offers recommendations to reduce distrust in artificial intelligence (AI)-driven digital forensics examinations. Secondly, they present a formal pre-concept for explainable digital forensics AI, along with various pertinent approaches for offering transparent explanations of AI models and their applicability in AI-based digital forensics analysis. The primary aim of this research is to examine different viewpoints on the clarity and interpretability of AI, with a specific focus on their implications for digital forensics and the evidence derived from AI algorithms.

3. PROBLEM

The technology and forensic tools used to examine digital evidence are constantly evolving, creating a dynamic landscape. Doyle[10] extensively researched quality management in forensic science and its correlation with bias and fairness. They concluded that a primary challenge across all forensic fields is the premature application of emerging scientific methods. Within this premature adoption of novel science, there are various factors that can lead to bias or inaccurate interpretations in software-driven digital forensics analysis. These factors encompass the design, input, model, and environment, with the most likely culprits being flawed algorithms/code and the software implementation of the algorithm in digital forensics tools. On the same note[8], pointed out that the general types of errors and biases in digital forensics arise from the tool's algorithms and their software implementation. These errors may manifest as exposure to irrelevant case information, base rate expectations from previous investigations, or the failure to assess alternative hypotheses.

However, the potential for bias and errors in these technologies and their decision-making processes has not been thoroughly examined. To address this issue, it is important to study the decision process of digital forensics and identify potential biases and errors. This knowledge can serve as the groundwork for transparency and the creation of effective strategies to reduce errors, ensuring they do not adversely affect the investigative process. Failure to detect and correct unreliable digital evidence poses a risk to the fair dispensation of justice and may lead to unjust convictions.

Two important factors in decision-making are reliability and bias. Reliability is the uniformity and ability to replicate decisions when confronted with the same evidence. Conversely, bias refers to the impact of task-irrelevant contextual information and other biases that may affect observations and conclusions. For instance, being aware of the suspect's race, confession status, or arrest history may shape expert judgments and decisions. The examination of bias involves assessing whether such information affects decision-making. Meanwhile, reliability can be measured by determining if decision-makers analyzing the same evidence with identical information bases consistently reach similar conclusions.

Bias can stem from various sources, such as contextual information. Contextual information can be categorized as either task-relevant or task-irrelevant for biasability[5]. In order to minimize contextual bias, it is essential to eliminate non-essential information from the decision-making process in forensics. While relevant information for the task is crucial, its introduction should be carefully timed and managed to prevent influencing the decision-making process with bias. Unfortunately, the digital forensics field presently lacks sufficient testing standards and protocols to assess the susceptibility of tools to bias during investigations. Although digital forensics offers decision-makers a dependable comprehension of digital traces, it cannot assure 100% accuracy of the tools employed to generate results in every instance.

3.1. Types of Bias

In this section, we describe in detail the different types of bias that can exist in software tools, as shown in Figure 1.

3.1.1. Algorithmic Bias

One of the main sources of bias in digital forensics tools is algorithmic bias, which occurs when the algorithms used in the tools produce discriminatory results. This can happen when the algorithms are trained on biased data sets or when the design of the algorithms themselves is flawed.

3.1.2. Data Bias

Another source of bias is data bias, which occurs when the data used in the forensic analysis is itself biased or incomplete. The impact of biased and unfair digital forensics tools can be significant, particularly in legal proceedings where digital evidence can be used to determine guilt or innocence. Biased evidence can lead to wrongful convictions or acquittals, resulting in serious injustices.

3.1.3. Sample Bias

Sample bias arises when the data utilized to train the algorithm does not accurately represent the domain for which the model is designed to be employed. In other words, sampling bias occurs when specific individuals in a group are consistently more prone to being selected for a sample than others. In the medical sciences, it is also known as ascertainment bias. Because sample bias jeopardizes external validity, particularly population validity, it restricts the generalizability of findings. That is, results from skewed samples can only be extrapolated to populations with similar traits. Autonomous vehicles and facial recognition software are two instances of sampling bias.

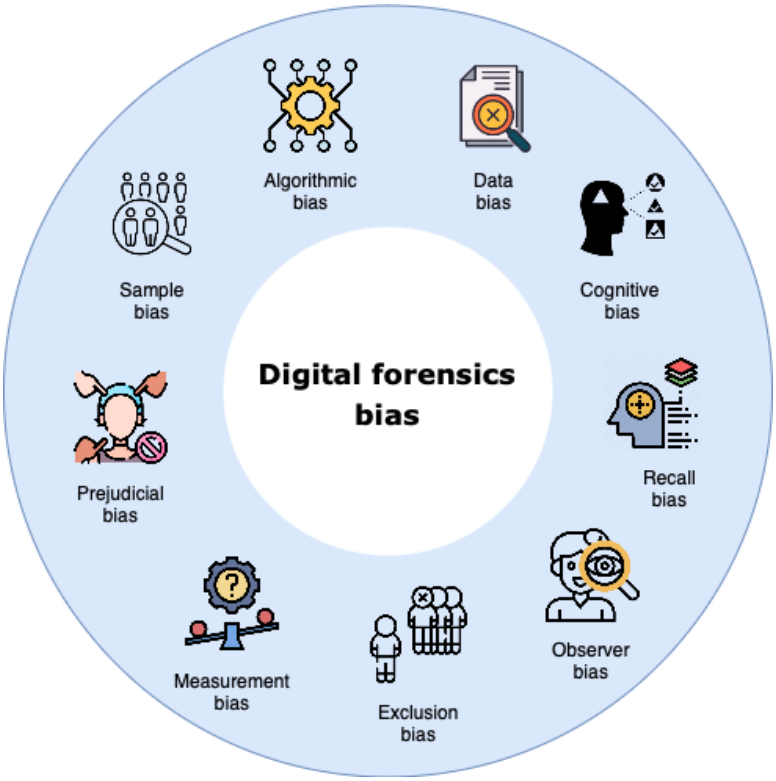

Figure 1. Types of bias in digital forensics investigations.

3.1.4. Prejudicial Bias

Prejudicial bias, also called racial bias, is the most popular among the biases. The data that is used to train the system is a reflection of the prejudices and assumptions of the creators and owners of the data.

3.1.5. Measurement Bias

Inconsistencies with the observation and/or measuring instrument lead to systematic value distortion. When measurements are made incorrectly, there is a bias that causes the data to change in a particular way. One example is the recognition of images[11].

3.1.6. Exclusion Bias

Exclusion bias occurs when a crucial data point is either absent or ignored from the data being used. In the stage of data preprocessing, this is also extremely typical. The removal of crucial data that was mistakenly thought to be irrelevant usually causes it to happen[11].

3.1.7. Observer Bias

Observer bias, sometimes referred to as "confirmation bias", occurs when an observer deliberately discovers the outcomes they anticipate seeing, regardless of what the evidence indicates. When researchers enter a study with preconceived notions based on their subjective experience from prior studies, this is known as observer bias. This also occurs when labelers influence their labeling task using their subjective expertise, resulting in flawed results[11].

3.1.8. Recall Bias

Recall bias has happened commonly in the data labeling phase. It occurs when similar types of data are labeled. The accuracy of the outcome is influenced by this[11]. Recall bias in software tools can lead to a reinforcement

of existing stereotypes or biases. If the training data for these tools is not sufficiently diverse or is skewed towards particular patterns or characteristics, the tool may inadvertently perpetuate and amplify these biases.

3.1.9. Cognitive Bias

Cognitive bias is the process that causes the human brain to simplify information processing through a filter of personal preferences or experiences. This process of simplification can lead to distorted perceptions, illogical reasoning, or judgments that are not based on a comprehensive evaluation of all available data.

4. SOFTWARE/AUTOMATION TOOLS DRIVEN DIGITAL FORENSICS ANALYSIS

In the digital context, interpreting the principle of presumption of innocence can be understood as a requirement to mitigate the influence of bias, uncertainties, and errors in the field of digital forensic science. However, judges and law enforcement officials often assume that digital evidence is reliable and trustworthy due to the increasing use of automated tools, which can create a false understanding that technology reduces errors and bias. This technological protection fallacy[12] disregards the numerous errors, biases, and uncertainties that can occur in digital investigations and can affect the presumption of innocence and fair trials.

AI- and Machine Learning (ML)-driven software and automation tools find widespread applications in our everyday activities, with diverse consequences for each sector. Yet, in domains where decisions hold substantial consequences for individuals or where accountability, transparency, or legal adherence is crucial — such as in health and law - there is a rising worry regarding the bias present in these AI systems[13]. As a result, there have been calls for audits of the application of AI-powered systems in different scenarios[14] to understand their behaviors.

With the progression of technology, the complexity of crimes facilitated by it also increases. This necessitates a transition from conventional forensic tools to more sophisticated and intelligent systems, such as AI, for detecting potential evidence. Nevertheless, there is notable uncertainty among courts, legal experts, and the public regarding the adoption of AI-based methods for extracting digital evidence. This skepticism is reasonable, considering the concerns raised about the transparency of closed-box AI software and their suitability for accuracy, bias, and reliability in the context of digital evidence extraction.

Over the past 20 years, machine-generated facts have largely replaced human fact-finding, resulting in increased accuracy[15]. Yet, there are substantial doubts regarding the legal status of digital evidence or findings generated by machines. This is particularly troubling since decisions based on scientific evidence can vary. Explainable AI (XAI)[16] is a research area to enhance the transparency of AI systems and the data they employ by exposing the operational components of the system, a concept often referred to as "glass-boxing"[17]. Given that decisions made by machines can have considerable implications for law enforcement and the entire criminal justice system, XAI becomes a pivotal area of attention. It endeavors to reduce or eliminate the obscurity of AI systems by dissecting intricate variables, all while maintaining a delicate equilibrium between transparency, performance, and accuracy.

To establish trust in a system, a mere accuracy evaluation is not sufficient. In some cases, accuracy alone may not accurately reflect real-world scenarios. Thus, in addition to accuracy, the bias in a system's decisions is crucial for determining the correctness of its outcome.

5. SCENARIOS OF BIAS IN DIGITAL FORENSIC CASES

Several real-world incidents have already highlighted the detrimental impact of bias in our society and digital forensic investigations. For our study and analysis, we selected some past incidents. Table 1 describes these

Table 1. Types of bias in each scenario

Scenario	Type of Bias
Scenario 1 (The Clearview Photo)	Sample, Prejudicial, Data
Scenario 2 (Now you See me)	Algorithmic
Scenario 3 (All About DNA)	Exclusion, Observer
Scenario 4 (A Tale of Two Petty Thieves)	Algorithmic, Prejudicial
Scenario 5 (Do not Neglect Me)	Prejudicial
Scenario 6 (Mapping Criminals)	Prejudicial, Data
Scenario 7 (Silence is Golden)	Observer, Algorithmic, Sample, Data

incidents and their corresponding biases.

5.1. Bias in facial recognition software

Law enforcement agencies utilize face recognition software to apprehend criminals. The main ethical problem, though, is prejudice, which underlies many of these requests. Face recognition algorithms of IBM, Microsoft, and Megvii (Face++) have been evaluated by MIT and Microsoft researchers under the "Gender Shades" project. According to their analysis, vulnerable gender misclassification has been on darker-skinned women, and its error rate is up to 34.4% than fair-skinned men[18]. It has been shown by researchers, including MIT's Joy Buolamwini, that technology frequently works better on males than women, better on white people than Black people, and worst of all on Black women. Significant biases still exist in some facial recognition software, despite some having improved in response. A 2018 study by the American Civil Liberties Union (ACLU) found that Amazon's facial recognition software misidentified people of color, particularly women. This raised concerns about the potential for bias in facial recognition software and the need for fairness in interpreting its results in criminal investigations[19].

5.1.1. Scenario 1 (The Clearview Photo)

A man from New Jersey named Nijeer Parks, who is the third known Black man to be mistakenly arrested using facial recognition technology, was charged with shoplifting and allegedly attempting to hit a police officer with a car. He was charged in February 2019 with the crimes at a Hampton Inn located in Woodbridge, New Jersey. A detective used a fake driver's license photo to perform a facial recognition search. The search yielded a match for Nijeer Parks. The detective compared Parks' state ID with the fake license and concluded they were of the same person. Despite Parks denying the resemblance and having been located 30 miles away from the incident when it occurred, he was identified by the police through facial recognition software called ClearviewAI. Mr. Parks was detained for ten days and had to pay approximately $5,000 for his legal defense. However, the case was eventually dismissed in November 2019 due to insufficient evidence[20].

5.1.2. Scenario 2 (Now you See me)

Five watches were stolen from a Shinola retail store in Detroit, and police were investigating who was responsible. According to authorities, the thief fled with goods worth an estimated 3,800 dollars. Security footage that had captured the event was retrieved by investigators. Detectives enlarged the blurry video and used face recognition software to identify the individual who appeared to be the culprit. As a result, they arrested Robert Julian-Borchak Williams in January 2020. Williams is the first known case in which a person was wrongly detained in the United States as a result of a false match generated by facial recognition technology. According to charge records seen by experts, facial recognition technology, used by Michigan State Police in a crime lab at the request of the Detroit Police Department, led to Williams' arrest. Despite Williams' and his attorneys' repeated assertions that the match produced by AI was inaccurate, Williams was pursued as a potential suspect. Studies from academia and the government have shown that white persons are misidentified by face recognition algorithms more frequently than people of color[21].

5.2. Inaccurate results in DNA analysis

In 2015, the FBI acknowledged that its DNA analysis software had produced inaccurate results in hundreds of cases over a period of years. This raised concerns about the potential for automated tools to produce flawed evidence and the need for fairness in interpreting their results in criminal trials.

5.2.1. Scenario 3 (All About DNA)

In December 2013, a group of Hasidic individuals assaulted a black student named Patterson in the Williamsburg area of Brooklyn, New York. Several of the attackers were associated with the Williamsburg Shomrim, a local neighborhood watch group for the Hasidic community. Patterson, who was intoxicated during the attack, was unable to identify his assailants because they all appeared to be young Hasidic men dressed similarly and with similar facial hair. As a result of the assault, Patterson suffered a severe eye injury, and one of his attackers even took one of his sneakers.

Six days after the attack, the stolen sneaker was found on a nearby roof. A portion of the heel, measuring three inches by six inches, was swabbed, revealing 97.9 picograms of DNA from at least two individuals, with Patterson's DNA being present. A picogram is one trillionth of a gram. Using in-house software, the laboratory determined that it was 133 times more likely than not that the remaining DNA belonged to Mayer Herskovic, a young father residing and working in Williamsburg with no prior criminal record.

Despite the absence of any other physical evidence tying Herskovic to the attack on Patterson, the trial judge convicted him of gang assault, resulting in a four-year prison sentence. There was no other inculpatory evidence apart from the DNA evidence. The victim and witnesses to the crime could not identify Herskovic at trial. Also, he was not positively identified on the surveillance video of the fleeing assailants. Furthermore, Herskovic was never part of the Shomrim, and he vehemently condemned the assault on Patterson.

On October 10, 2018, the conviction was overturned, and the charges were dismissed by the Appellate Division for the Second Judicial Department of the Supreme Court of the State of New York[22].

5.3. Bias in Recidivism Software

In 2019, ProPublica published an investigation into using proprietary software called "Compas" in the criminal justice system. The software is used to predict the likelihood of defendants reoffending, but the investigation found that the algorithm was biased against black defendants, resulting in them being incorrectly labeled as higher risk than white defendants with similar backgrounds[23].

5.3.1. Scenario 4 (A Tale of Two Petty Thieves)

Brisha Borden, an 18-year-old black woman, and her friend stole a child's bike and scooter for a ride down the street in Coral Springs, a Fort Lauderdale suburb, with the intention of picking up her god-sister from school. They dropped the items after the child's mother accosted them, leading to their arrest on charges of burglary and petty theft for the items, which were collectively valued at $80. One year earlier, Vernon Prater, a 41-year-old white man, was arrested for shoplifting tools worth $86.35 from a nearby Home Depot store.

Prior to their respective arrests, Prater had already been convicted of armed robbery and attempted armed robbery, serving five years in prison for those offenses, in addition to facing another armed robbery charge. In contrast, Borden had misdemeanor charges as a juvenile. Despite Prater's extensive criminal history, a computer program called "Compas" predicted that Prater had a low risk of committing future crimes, while it indicated a high-risk assessment for Borden.

However, it is worth noting that this prediction turned out to be inaccurate. Prater is currently serving an eight-year prison sentence for a subsequent break-in at a warehouse and theft of electronics, while Borden has

not faced any new criminal charges.[22].

5.4. Bias in child exploitation material

In a 2020 study, researchers at the University of California, Irvine analyzed the use of automated tools for identifying child exploitation material (CEM) in digital devices. The study found that the tools had a high false positive rate, resulting in innocent individuals being falsely accused of possessing CEM. The study also found that the tools were less accurate when analyzing images of people from non-Western countries, indicating potential bias in the algorithms[24].

5.4.1. Scenario 5 (Do not Neglect Me)

Child welfare authorities in Oregon have opted to discontinue the use of an algorithm for assessing which families should be subject to investigations by social workers. Instead, they are adopting a new procedure designed to improve decision-making with a focus on promoting racial equity. This decision was prompted by an Associated Press investigation that raised concerns about the transparency, dependability, and racial imbalances associated with similar algorithmic tools used in Pennsylvania, California, and Colorado. These tools disproportionately flagged Black children for mandatory neglect investigations, leading to a reevaluation of their implementation. Oregon's Department of Human Services has stated that the algorithm will no longer be used to reduce disparities and bias in child abuse and neglect investigations[25].

5.5. Bias in Automated Systems

Bias in automated systems is a pressing concern with regard to digital forensic investigations. These biases can originate from various sources, including data collection, algorithms, and even human biases embedded in the design and interpretation of digital forensic tools. Such biases can lead to inaccuracies in evidence analysis, potentially compromising the integrity of investigations and court proceedings.

5.5.1. Scenario 6 (Mapping Criminals)

PredPol's algorithm creates a heat map using reported crimes and arrests rather than real crime scenes. Therefore, its predictions may come true. A feedback loop occurs when officers are sent to areas where there are already a high number of arrests being made by the police. The maps would be comparable in an ideal world. As an example, in the state of California, instead of focusing on the areas where drug criminality really occurred, PredPol instead led the police toward predominantly Black communities such as West Oakland and International Boulevard. Even if white individuals use illicit substances at a greater rate than minorities, mostly white areas, such as Rockridge and Piedmont, received a pass. Wherever violence occurs, there are nearly 200 times more drug arrests in largely black communities in Oakland than in other areas[26].

5.5.2. Scenario 7 (Silence is Golden)

In August 2021, Micheal Williams was arrested and charged with the murder of a local man who had asked for a ride during a night of protests against police brutality in May. The primary evidence against Williams came from the security footage showing a car passing through an intersection and a loud bang picked up by surveillance microphones. Prosecutors claimed that ShotSpotter technology, which analyzed sounds detected by the sensors, indicated that Williams was responsible for the shooting. For almost a year, Williams was incarcerated before the case against him was dismissed due to a lack of evidence[27].

On the same line, an investigation conducted by the Associated Press revealed significant issues with using ShotSpotter as evidence in criminal trials. ShotSpotter, a technology that utilizes proprietary algorithms to detect gunfire, can miss live gunfire or mistake other sounds, such as fireworks or car backfires, for gunshots. The company's algorithms are kept confidential, creating a lack of transparency in how the technology works. Moreover, ShotSpotter employees can alter the source of sounds picked up by the sensors, potentially introducing human bias into the algorithm[27].

6. RECOMMENDATIONS FOR MITIGATING BIAS IN DIGITAL FORENSIC TOOLS

The development and application of digital forensic tools in investigative processes have become increasingly critical. These tools, however, may be biased, affecting the accuracy and fairness of investigations and legal proceedings. In response to this pressing issue, we present a set of comprehensive recommendations aimed at mitigating bias in digital forensic tools. These recommendations encompass diverse aspects, including the careful curation of datasets, thorough validation processes, the need to contextualize scenarios, the formulation of hypotheses, establishing frameworks for tool testing, and adopting best practices for interpreting results. By addressing these key dimensions, we aim to enhance the reliability, objectivity, and equity of digital forensic practices, ultimately contributing to the integrity of the criminal justice system.

6.1. Dataset Analysis

Analytical inaccuracies may arise in machine-generated outcomes if the machines modify their operational parameters in unforeseen manners. This problem might stem from training data with a limited number of samples, which may not accurately reflect real-world scenarios or be inadequate for drawing inferences about future observations. Incorporating an excessive number of variables into the model may also lead to the model being trained to acquire irrational representations. For instance, a surveillance system's predictive crime detection algorithm monitors criminal activities and notifies officers either before or precisely when a crime occurs. According to reports, through the analysis of crime-related instances from surveillance camera data, the algorithm acquired the ability to identify three consecutive handshakes as probable indicators of narcotic transactions. Although this line of reasoning might appear reasonable, it could fail to identify instances of drug-related activities in the actual world if such a pattern is nonexistent.[28]. Using such examples in a court case to argue against the reliability of software and machine drive methods would only further intensify public skepticism toward evidence generated by machines. Although software-generated decisions can be affected by algorithmic biases, these biases are often linked more to the training data used than to the technical aspects of data processing.

Additionally, forensic examinations begin with a potential evidence source of unknown content, implying that the inputs to the entire forensic procedure are also unknown[29]. To guarantee precision and dependability in digital forensic analysis, it is essential for the digital forensics community to uphold a reference database covering the full range of conditions expected during the analysis.

6.2. Comprehensive Validation

During digital forensic analysis, errors and uncertainties can arise due to the algorithms and software implementation of the tools used. These mistakes can hinder the effective functioning of the tools and result in deficiencies in the analysis. These uncertainties are undetected flaws in the algorithms, software bugs, hardware limitations and flaws, and other similar problems. To reduce uncertainties, researchers can improve tool validation by incorporating test cases that are prone to revealing defects in the algorithms and software, potentially inducing conditions that could lead to tool malfunction[7]. Hence, the digital forensics community needs to compile reference data that accurately mirrors the entire spectrum of conditions anticipated in digital forensic analysis.

Designing and creating test cases that can comprehensively assess the functionality of a tool is challenging in the absence of insights into the code and algorithm design[30]. Generating and maintaining detailed and documented test datasets is a primary element and barrier in the testing process within the field of digital forensics. These datasets are crucial in attempting to thoroughly test a tool's functionality[31]. To thoroughly test a tool's functionality, the dataset used must contain comprehensive data that can exhaustively test the tool. The dataset that is generated should be thoroughly documented and should contain evidence to assess all aspects and levels of interpretation of the tool, along with its complete set of features. Developing a framework that outlines the structure for dataset creation and its utilization in formulating and executing tool testing has

the potential to enhance the accessibility of testing for practitioners.

Breaking down forensic responsibilities into smaller sub-tasks, specifying each test case to mirror potential conditions encountered in analysis, and creating a process to incorporate these test cases into a testing system could prove advantageous[3]. By formulating an exhaustive compilation of specifications and test cases, it becomes feasible to evaluate the suitability of a tool for a particular objective.

Formal validation is essential to uncover tool bias. Neglecting to recognize the limitations and errors of the tools may lead to the reopening of prior cases for reassessment upon their detection. Given that a single tool or method may be employed in numerous investigations and trials, it is vital to recognize limitations and errors to prevent potential adverse impacts on the presumption of innocence and the overall fairness of the trial.

6.3. Contextualize the scenario

The findings regarding bias highlight the importance of researching contextual bias and the management of context. It is essential to enhance our understanding of which information holds the most significant potential to influence decisions and to differentiate between contextual information that is relevant to the task and that which is not. Equipped with this knowledge, digital forensics context management methods can be developed that grant access to task-relevant contextual data while restricting or prohibiting access to task-irrelevant contextual data.

In digital forensics, the same tools are often used for different purposes, such as terrorism prevention, child pornography detection, and murder investigations. These tools may use techniques such as sentiment analysis, link and text analysis, and others. However, questions have been raised regarding the algorithms' fairness and bias in these tools. The precision of an algorithm relies on previous problem-specific knowledge to attain performance superior to random chance. This means that the algorithm selection, accurate interpretation, and assumptions about the input data based on knowledge specific to the type of investigation, such as child pornography, murder, or terrorism cases, will determine how accurate a digital forensic analysis is[6].

Forensic tools designed for general use are not tailored for specific domains of investigation and do not explicitly state the assumptions about the characteristics of the dataset. They employ a set algorithm implementation, and if they are closed-source, verifying the algorithm and feature selection process is only feasible if the source code is revealed. In cases related to child pornography, specific methods are utilized to ascertain the identity of the victim and understand how the illegal material is handled and stored by the suspect. These methods involve automated comparisons and searches within databases. Additional methods concentrate on identifying skin or investigating messaging platforms to detect conversations related to grooming and child pornography.

It is critical to evaluate each case and its possible implications before incorporating automation models into digital forensics. Various cases may necessitate distinct methods and interpretation criteria. For instance, examining emails for indications of deliberate deletions to hide potentially incriminating activities may entail distinct requirements for interpretation compared to establishing accountability in electronic contract agreements conducted via emails with involvement from multiple parties. Being aware of the context can provide a more comprehensive understanding of the investigation's scope.

6.4. Creation of hypothesis

During a digital forensic investigation, both pertinent and non-pertinent information is abundant in digital evidence, and it is challenging to selectively exclude them. A potential bias mitigation technique to combat this would be to begin with a fair set of alternative hypotheses that reflect both guilt and innocence, methodically taking into account each hypothesis as the study progresses, and openly communicating the findings on each. It is crucial to disclose task-relevant and task-irrelevant information that guides the examination to facilitate

review and detection of biased decision-making.

6.5. Framework for tool testing

Designing a framework/blueprint to enable tool testing and validation for bias and fairness could establish a manageable standard, leading to the standardization of tool testing in digital forensics and increasing confidence in the generated results. It can also be used by tool developers for effective guidance on creating and developing new tools while keeping in mind the best practices to ensure that forensic tools are free from bias.

6.6. Understanding different kinds of artifacts

It is crucial to understand the different kinds of artifacts present in digital forensic investigations. When it comes to generating artifacts, such as identifying whether a contraband image was downloaded manually from the internet or through some other method, a thorough understanding of artifacts could offer approximations on the confidence levels of opposing hypotheses. By establishing a model for artifacts linked to various use case scenarios (such as passive downloads through banner ads), it becomes possible to quantify the probability of each hypothesis using evidence gathered from the targeted device. Through a more thorough understanding of what ought to be seen during a given use case scenario, the correlations generated by digital forensic tools can be verified, guaranteeing that they are free from bias. Furthermore, comprehending the connections between artifacts could improve the capacity to systematically produce realistic reference data.

6.7. Recommendations for machine learning models

6.7.1. Preprocessing

Digital forensics investigations involve examining diverse, unstructured, and unorganized digital artifacts. Thus, before these variables are input into a ML model, they must undergo preprocessing.

6.7.2. Feature Importance

After finishing the preprocessing stage, it is crucial to assess and analyze the influence, importance, and significance of each training variable on the model's predictions. This step is extremely important to remove context-unrelated variables that misguide the model.

6.7.3. Visualization of the model

Visual explanation, which is particularly effective and common in model-specific approaches. In this approach, feature importance techniques are frequently used to provide explanations. While feature importance is beneficial, visualization approaches offer an innovative way to physically observe the interaction of influential variables during the process which can further help determine the source of bias in the software.

6.7.4. Text explanation

The incorporation of plain natural language explanations to closed-box models is an underexplored approach in the literature [9]. Descriptions in the text of each decision-making component of a model can be provided. Sometimes, the text explanations are presented in a rule-based style (if ... then), in which all decision-making components are semantically explained. Integrating this method with other techniques, such as feature importance and visualization, can prove to be highly beneficial in the development of software. This contributes to constraining the bias and fostering fairness in software and automation tools employed in digital forensic examinations.

6.7.5. Explainable Artificial Intelligence

In AI, XAI refers to techniques used to analyze the decision-making process of ML models. XAI techniques can be used to detect and correct biases in ML models [32]. By utilizing these techniques, we can gain insights into the inner workings of ML models, thus enabling us to identify the factors that drive their predictions. Through this understanding, we can eliminate bias and improve the fairness and accuracy of the models. By providing

Table 2. Recommendations to mitigate bias in each scenario

Scenario	Type of Bias	Recommendations
Scenario 1	Sample, Prejudicial, Data	• Expand dataset with more inclusive data • Comprehensive validation needs to be performed through different test cases
Scenario 2	Algorithmic	• Trained dataset needs to be tested on several test cases • Data labeling needs to be corrected
Scenario 3	Exclusion, Observer	• Other pieces of evidence need to be considered • Law enforcement officers need to know DNA testing can be fallible
Scenario 4	Algorithmic, Prejudicial	• Algorithm needs to consider past offenses of the person • Prejudices about race and gender should be eliminated
Scenario 5	Prejudicial	• Comprehensive validation should be performed through different test cases • There should be context awareness in the choice of tool
Scenario 6	Prejudicial, Data	• The dataset can be expanded to include more inclusive data • The scenario should be contextualized while choosing a tool
Scenario 7	Observer, Algorithmic, Sample, Data	• Comprehensive validation needs to be performed using different test cases • The dataset should be expanded to include more samples • Extensive validation needs to be performed through different test cases

transparency and interpretability into the decision-making processes of AI models, XAI allows investigators and forensic analysts to understand how predictions and conclusions are reached. This transparency is pivotal in identifying and rectifying biases that may exist in the data sources, training datasets, or algorithms. Furthermore, XAI can be used as a tool to reduce bias by enabling continuous monitoring and auditing of software tools in digital forensics.

6.8. Interpretation of Results

In the process of scrutinizing digital evidence, digital forensics examiners utilize their expertise to interpret and extract meaning from the observed traces. These interpretations are then conveyed through analysis reports or expressed as expert testimony. These interpretations underpin the conclusion of the case. Sunde *et al.*[5] disclose that when the examiner holds the belief that the suspect is innocent, they are inclined to discover fewer traces of evidence. Conversely, if the digital forensics examiner believes the suspect is guilty, they tend to identify more traces of evidence. Thus, examiners should strive to maintain objectivity and avoid bias to ensure their findings are accurate and reliable by conducting blind analysis, avoiding assumptions, using a standard methodology, documenting every detail of the case, and verifying their findings.

6.9. Blockchain Application

Blockchain can reduce bias and enhance fairness in software automation by providing a transparent and tamper-resistant ledger used for the management, storage, sharing, and retrieval of data[33]. Blockchain technology ensures that all relevant parties have access to the same information, eliminating biases that may arise from selective disclosure or manipulation of data. Additionally, the use of smart contracts in blockchain-based automation can enforce predefined rules and procedures without human bias or interference. Furthermore, the decentralized nature of blockchain technology allows for a more inclusive and diverse network of participants. This reduces the risk of bias and ensures that decisions and outcomes are not influenced by a single entity or controlling authority. Using the blockchain for automation in digital forensics can lead to reduced bias and increased fairness by providing transparency, traceability, and immutability in systems that ensure equal opportunities[34]. Moreover, by reducing human involvement and relying on automated procedures, the risk of bias stemming from personal judgments or prejudices is minimized.

7. DISCUSSION AND CONCLUSION

In this research, we studied and examined the different kinds of biases that exist in software and automation tools that are used by law enforcement agencies and in court cases. We analyze these biases by studying different scenarios and mapping the varying bias types, as shown in Table 1.

We analyze the scenarios in Section 4 and map the corresponding recommendations from Section 5. Table 2

describes our analysis and results. It outlines specific types of biases, such as Sample, Prejudicial, Data, Algorithmic, Exclusion, and Observer biases, and pairs them with our tailored recommendations to address each unique challenge. This can aid us in understanding the multifaceted nature of bias and the diverse strategies required for effective mitigation.

First, the mitigation involves expanding and diversifying datasets to combat bias inherent in data sources and algorithms. In addition, in scenarios dealing with prejudicial or data biases, expanding the dataset to include more inclusive data is a recommended strategy. This is coupled with the need for comprehensive validation across different test cases, ensuring that the solutions are robust and widely applicable. In cases of algorithmic bias, the emphasis shifts to scrutinizing and correcting the training datasets and data labeling processes. Moreover, the table highlights the necessity of context awareness and the consideration of additional evidence in scenarios where observer or exclusion biases are prevalent. By integrating these recommendations, organizations and individuals can develop more equitable and unbiased systems and processes, reflecting a deep understanding of the complexities of bias and its impact.

In conclusion, this paper has shed light on the critical issue of bias in software tools employed within the field of digital forensics. By examining various scenarios and their potential implications, we have underscored the urgency of addressing bias to ensure the integrity and fairness of investigations and legal proceedings. Through the development of a comprehensive framework encompassing key elements such as dataset curation, validation processes, contextualization, hypothesis formulation, testing methodologies, and result interpretation, we have provided a roadmap for practitioners and researchers in the field to mitigate bias effectively.

Recognizing bias in digital forensic tools is not merely a theoretical concern; it has real-world implications that can affect individuals' lives and the credibility of the criminal justice system. As technology continues to advance, the imperative to scrutinize and rectify bias becomes even more crucial. The recommendations presented in this paper serve as a starting point for fostering greater transparency, accountability, and equity in digital forensics. By implementing these measures and continuously refining our methodologies, we can strive for a future where digital forensic tools not only facilitate thorough investigations but also uphold the principles of justice and fairness upon which our legal system is built. We hope this paper sparks further research and discussion on this vital topic, ultimately contributing to more reliable and equitable digital forensic practices. In future work, we intend to implement and test some of the recommended tools and processes on real-life tools and software.

DECLARATIONS

Authors' contributions

Made substantial contributions to the conception and design of the study and performed data analysis and interpretation: Jinad R, Gupta K, Simsek E

Performed data acquisition and provided administrative, technical, and material support: Zhou B

Availability of data and materials

Not applicable.

Financial support and sponsorship

None.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2024.

REFERENCES

1. Council TUNPC. *Digital forensic science strategy*. Available from: https://npcc.police.uk/Digital%20Forensic%20Science%20Strategy%2020\20.pdf. [Last accessed on 19 Jan 2024]
2. Carrier B, Spafford EH. Getting physical with the digital investigation process. Available from: https://www.utica.edu/academic/institutes/ecii/publications/articles/A0AC5A7A-FB6C-325D-BF515A44FDEE7459.pdf[Last accessed on 19 Jan 2024]
3. Guo Y, Slay J, Beckett J. Validation and verification of computer forensic software tools—Searching Function. *Digit Invest* 2009;6:S12–22. DOI
4. Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine* 2017;38:50–7. DOI
5. Sunde N, Dror IE. A hierarchy of expert performance (HEP) applied to digital forensics: Reliability and biasability in digital forensics decision making. *Forens Sci Int-Digit* 2021;37:301175. DOI
6. Stoykova R. Digital evidence: Unaddressed threats to fairness and the presumption of innocence. *Comput Law Secur Rev* 2021;42:105575. DOI
7. Hughes N, Karabiyik U. Towards reliable digital forensics investigations through measurement science. *WIREs Forensic Science*2020;2:e1367. DOI
8. Horsman G. Tool testing and reliability issues in the field of digital forensics. *Digit Invest* 2019;28:163–75. DOI
9. Solanke AA. Explainable digital forensics AI: Towards mitigating distrust in AI-based digital forensics analysis using interpretable models. *Forensic Science International: Digit Invest* 2022;42:301403. DOI
10. Doyle S. Quality management in forensic science. Academic Press; 2018.p.1-387. Available from: https://www.sciencedirect.com/book/9780128054161/quality-management-in-forensic-science?via=ihub= [Last accessed on 19 Jan 2024]
11. Chakraborty K. Fairness in machine learning: eliminating data bias Available from: https://www.techopedia.com/fairness-in-machine-learning-eliminating-data-bias/2/34389.[Last accessed on 19 Jan 2024]
12. Dror IE. Cognitive and human factors in expert decision making: six fallacies and the eight sources of bias. *Anal Chem* 2020;92:7998–8004. DOI
13. Coyle D, Weller A. "Explaining" machine learning reveals policy challenges. *Science* 2020;368:1433–4. DOI
14. Schneider J, Breitinger F. Towards Ai forensics: did the artificial intelligence system do it? *J Inf Secur Appl* 2020;76:103517 DOI
15. Roth A. Trial by machine. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2743800.[Last accessed on 19 Jan 2024]
16. Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR. Explainable AI: interpreting, explaining and visualizing deep learning. Springer Nature; 2019. p. XI, 439. DOI
17. Gross-Brown R, Ficek M, Agundez JL, Dressler P, Laoutaris N. Data transparency lab kick off workshop (DTL 2014) report. *SIGCOMM Comput Commun Rev* 2015;45:44-8. DOI
18. Zeng Y, Lu E, Sun Y, Tian R. Responsible facial recognition and beyond. *arXiv preprint arXiv:1909.12935* 2019. DOI
19. Snow J. Amazon's face recognition falsely matched 28 members of congress with mugshots. Available from: https://www.aclu.org/news/privacy-technology/amazons-face-recognition-falsely-matched-28.[Last accessed on 19 Jan 2024]
20. Hill K. Another arrest, and jail time, due to a bad facial recognition match. Available from: https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html. [Last accessed on 19 Jan 2024]
21. Allyn B. "the computer got it wrong": How facial recognition led to false arrest of Black Man. Available from: https://www.npr.org/2020/06/24/882683463/the-computer-got-it-wrong-how-facial-recognition-led-to-a-false-arrest-in-michig. [Last accessed on 19 Jan 2024]
22. Kirchner L. Thousands of criminal cases in New York relied on disputed DNA testing techniques. Available from: https://www.propublica.org/article/thousands-of-criminal-cases-in-new-york-relied-on-disputed-dna-testing-techniques.[Last accessed on 19 Jan 2024]
23. Lee NT, Resnick P, Barton G. Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. Available from: https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/. [Last accessed on 19 Jan 2024]
24. Irvine C. NIH awards over $2.5 million for research to improve interviewing of young witnesses. Available from: https://news.uci.edu/2021/09/30/nih-awards-over-2-5-million-for-research-to-improve-interviewing-of-young-witnesses/. [Last accessed on 19 Jan 2024]
25. Burke SH, Garance. Oregon dropping controversial algorithm used to trigger investigations in child abuse cases. Available from: https://www.statesmanjournal.com/story/news/local/oregon/2022/06/02/\oregon-drops-ai-algorithm-tool-for-child-abuse-cases-investigations-over-race-bias-concerns/65358612007/. [Last accessed on 19 Jan 2024]
26. IV JS. crime-prediction tool PredPol amplifies racially biased policing, study shows. Available from: https://www.mic.com/articles/156

286/crime-prediction-tool-pred-pol-only-amplifies-racially-biased-policing-study-shows. [Last accessed on 19 Jan 2024]
27. Garance Burke MM. How ai-powered tech landed man in jail with scant evidence. Available from: https://apnews.com/article/artificial-intelligence-algorithm-technology-police-crime-7e3345485aa668c97606d4b54f9b6220. [Last accessed on 19 Jan 2024]
28. Roth A. Machine testimony. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2893755. [Last accessed on 19 Jan 2024]
29. Marshall AM, Paige R. Requirements in digital forensics method definition: Observations from a UK study. *Digit Invest* 2018;27:23–9. DOI
30. Glisson WB, Storer T, Buchanan-Wollaston J. An empirical comparison of data recovered from mobile forensic toolkits. *Digit Invest* 2013;10:44–55. DOI
31. Grajeda C, Breitinger F, Baggili I. Availability of datasets for digital forensics–and what is missing. *Digit Invest* 2017;22:S94–105. DOI
32. Gerlings J, Shollo A, Constantiou I. Reviewing the need for explainable artificial intelligence (xAI). *arXiv preprint arXiv:201201007* 2020. DOI
33. Sobreira Leite G, Bessa Albuquerque A, Rogerio Pinheiro P. Process automation and blockchain in intelligence and investigation units: an approach. *Applied Sciences* 2020;10:3677. DOI
34. Zhang P, Ding S, Zhao Q. Exploiting blockchain to make AI trustworthy: a software development lifecycle view. *ACM Comput Surv* 2023. DOI