Fair, Unbiased, and Interpretable Algorithms in Digital Humanism

Rahulrajan Karthikeyan[1][0009−0001−7691−5384] and
Moses Boudourides[2][0000−0002−6157−5647]

[1] Arizona State University, Tempe, AZ, USA
rkarthi5@asu.edu
[2] Northwestern University, Evanston, IL, USA
Moses.Boudourides@northwestern.edu

Abstract. As algorithmic systems become increasingly integral to decision-making in sensitive domains such as criminal justice and healthcare, issues of fairness, transparency, and accountability have become paramount. This chapter examines through the lens of digital humanism and explores how predictive models can perpetuate societal inequities, using the COMPAS algorithm as a central case study in the criminal justice system. Empirical evidence has shown that COMPAS exhibits racial bias, disproportionately misclassifying Black defendants as high-risk. Through experimental analysis using Random Forest models and adversarial debiasing techniques supported by tools like AI Fairness 360 and LIME, this demonstrates how fairness-aware machine learning can mitigate algorithmic bias without significantly compromising accuracy. We extend this inquiry into the healthcare domain, showcasing how adversarial debiasing improves equity in diagnostic algorithms, particularly during the COVID-19 pandemic. Further, the chapter interrogates the role of interpretability and explainability in forensic science, highlighting the dangers of opaque methodologies, cognitive bias, and the prosecutor's fallacy. Case examples from fingerprint analysis and probabilistic evidence underscore the need for transparent, interpretable, and bias-aware forensic tools such as AFIS. Collectively, the findings underscore the urgent need for technically robust and ethically grounded approaches to algorithmic design, capable of delivering not just accurate predictions but justice-aligned outcomes.

Keywords: Algorithmic Fairness, COMPAS, Interpretability, Explainability, Forensic Science, Adversarial Debiasing, Digital Humanism

1 Introduction

The proliferation of algorithmic decision-making systems across critical domains—from criminal justice to healthcare—has brought unprecedented efficiency and scale to complex societal challenges. However, this technological advancement has also exposed fundamental questions about fairness, accountability, and the preservation of human dignity in automated processes. Recent

studies have demonstrated that algorithmic systems can perpetuate and amplify existing societal biases, leading to discriminatory outcomes that disproportionately affect marginalized communities [8, 13, 2, 21].

For example, in the U.S. criminal justice system, the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm has been shown to misclassify Black defendants as high-risk at significantly higher rates than their white counterparts. These findings have brought issues of fairness and transparency in black-box decision-making systems to the forefront of academic and public discourse.

This work examines these challenges through the lens of Digital Humanism, an interdisciplinary approach that seeks to ensure technological development serves human flourishing while respecting fundamental rights and values. Digital Humanism, as articulated in the Vienna Manifesto, emphasizes the need for technology that is transparent, accountable, and aligned with human values [27].

1.1 Digital Humanism Framework

Digital Humanism represents a paradigm shift in how we approach the development and deployment of algorithmic systems. Rather than viewing technology as value-neutral, Digital Humanism recognizes that all technological artifacts embody certain values and assumptions about the world. This framework demands that we explicitly consider the human and social implications of our technological choices. As Tim Berners-Lee, the founder of the World Wide Web, has observed, "the system is failing," pointing to the monopolization of digital platforms, the rise of extremist content, the formation of filter bubbles, and the erosion of privacy as evidence that technology is not automatically serving human interests [5]. Digital Humanism represents a response to these challenges, offering both a critique of current technological practices and a vision for how technology can be developed and deployed in ways that promote human flourishing.

The intellectual roots of Digital Humanism can be traced to the humanistic traditions of the Renaissance and the Enlightenment, which emphasized human dignity, rational inquiry, and the potential for human progress through knowledge and ethical action [28]. However, Digital Humanism is not simply a nostalgic return to earlier humanistic ideals but rather a critical engagement with these traditions in light of contemporary technological challenges. It recognizes that the digital age requires new forms of humanistic thinking that can grapple with the unprecedented scale and complexity of technological systems while maintaining commitment to core humanistic values.

Central to Digital Humanism is the recognition that technology is never neutral but always embodies particular values, assumptions, and power relations [30]. As Langdon Winner famously argued, "artifacts have politics," meaning that technological systems inevitably shape social relations and distribute power in particular ways [31]. Digital Humanism extends this insight by arguing that if technology inevitably embodies values, then we have a responsibility to ensure that these values align with human dignity and democratic principles rather than serving narrow commercial or political interests.

This perspective challenges the widespread assumption that technological development is an autonomous process driven by purely technical considerations. Instead, Digital Humanism insists that technological development is fundamentally a social and political process that reflects the values and interests of those who design, fund, and deploy technological systems [15]. This recognition opens up space for democratic participation in technological development and creates opportunities for ensuring that technology serves broader human interests rather than narrow elite concerns.

The Vienna Manifesto on Digital Humanism, signed by over 1000 leaders worldwide, articulates this vision by calling for technology that promotes democracy and inclusion, protects privacy and freedom of speech, ensures accountability and transparency, and maintains human agency in decision-making processes [27]. These principles reflect a commitment to what we might call "technological democracy"—the idea that technological development should be subject to democratic oversight and should serve democratic values.

Building on the work of Mark Coeckelbergh and other leading scholars in the field, we can identify a number of core components that define Digital Humanism as both a theoretical framework and a practical approach to technological development [10]. These components provide a comprehensive framework for understanding how humanistic principles can be integrated into technological practice.

Digital Humanism begins with a fundamental commitment to defending human dignity against tendencies to reduce humans to mere data points or algorithmic inputs. This component recognizes that contemporary technological systems often embody what Coeckelbergh calls a "negative anthropology"—defining humans primarily in opposition to machines rather than in terms of their positive capacities for creativity, moral reasoning, and social connection [9]. Digital Humanism challenges this reductive view by insisting on a richer understanding of human nature that recognizes the complexity, creativity, and moral agency that distinguish human beings from algorithmic systems.

This principle has profound implications for algorithmic design. Rather than treating humans as passive subjects to be optimized or predicted, Digital Humanism demands that algorithmic systems be designed to enhance human agency and support human flourishing. This means, for example, that predictive algorithms should be designed not merely to maximize accuracy but to provide information that helps humans make better decisions while preserving their autonomy and dignity.

Another component of Digital Humanism emphasizes the importance of maintaining human control over technological systems, particularly in domains where algorithmic decisions can significantly impact human lives. This principle responds to growing concerns about the "black box" nature of many AI systems and the tendency for algorithmic automation to displace human judgment in critical domains [23]. Digital Humanism insists that humans must retain meaningful control over technological systems, particularly in high-stakes contexts such as criminal justice, healthcare, and education.

This does not mean rejecting automation entirely but rather ensuring that automated systems are designed to support and enhance human decision-making rather than replace it. In the context of algorithmic fairness, this principle suggests that bias mitigation techniques should be designed not only to improve statistical measures of fairness but also to provide human decision-makers with the information and tools they need to make just and informed decisions.

Digital Humanism demands that technological systems be explicitly designed to embody and promote human values such as integrity, equality, justice, and democratic participation [14]. This goes beyond traditional approaches to technology ethics that focus primarily on avoiding harm to actively promoting positive human values through technological design. This principle aligns with emerging approaches such as value-sensitive design and responsible innovation but places these approaches within a broader humanistic framework that emphasizes the social and political dimensions of technological development.

In the context of algorithmic fairness, this principle suggests that bias mitigation techniques should be understood not merely as technical optimizations but as implementations of broader commitments to justice and equality. This perspective can help guide the development of fairness metrics and bias mitigation techniques by grounding them in explicit ethical and political commitments rather than treating them as purely technical problems.

Digital Humanism recognizes that the challenges posed by contemporary technology cannot be adequately addressed by any single discipline but require collaboration across technical and humanistic fields [29]. This principle reflects the understanding that technological problems are simultaneously technical, social, ethical, and political problems that require diverse forms of expertise to address effectively.

The Digital Humanist interdisciplinary approach has important implications for how we approach algorithmic fairness. Rather than treating bias as a purely technical problem to be solved through mathematical optimization, Digital Humanism suggests that addressing algorithmic bias requires collaboration between computer scientists, ethicists, legal scholars, social scientists, and affected communities. This collaborative approach can help ensure that technical solutions are grounded in broader understanding of social justice and human rights.

Digital Humanism emphasizes the importance of community engagement and democratic participation in technological development [25]. This principle recognizes that technological systems inevitably affect entire communities and that those who are affected by technology should have a voice in how it is developed and deployed. This goes beyond traditional approaches to technology assessment that focus primarily on expert evaluation to include meaningful participation by affected communities.

In the context of algorithmic fairness, this principle suggests that bias mitigation efforts should include meaningful participation by communities that are affected by algorithmic systems. This might involve community-based participatory research, public engagement processes, or other mechanisms for ensuring

that affected communities have a voice in how algorithmic systems are designed and evaluated.

Similarly, the interpretability challenges in forensic science and other disciplines reflect a broader need for transparency and accountability in expert systems. The Digital Humanism framework provides a valuable lens for understanding these challenges and developing solutions that prioritize human welfare and justice.

This weaves from real-world examples—including the COMPAS risk assessment tool and healthcare diagnostics—to explore how fairness-aware algorithms, adversarial debiasing, and interpretability frameworks can mitigate systemic bias. By situating these technical strategies within the principles of Digital Humanism, we emphasize that fairness, transparency, and accountability must be embedded into the design and governance of predictive systems. Ultimately, the chapter makes the case for algorithms that are not only accurate, but also fair, unbiased, and interpretable—serving human values rather than undermining them.

2 Addressing Algorithmic Bias through Digital Humanism

The challenge of algorithmic bias represents one of the most pressing concerns in contemporary AI ethics. Bias in machine learning systems can arise from multiple sources: biased training data, flawed model assumptions, or discriminatory feature selection. These biases can have profound real-world consequences, particularly in high-stakes domains like criminal justice and healthcare.

From a Digital Humanism perspective, addressing algorithmic bias is not merely a technical challenge but a moral imperative. It requires us to consider not just the accuracy of our models but their impact on human dignity and social justice. This section explores various approaches to bias mitigation, with particular attention to adversarial debiasing techniques.

2.1 Fairness through Adversarial Debiasing

Adversarial debiasing represents a promising approach to creating fairer machine learning models. By training a model to make accurate predictions while simultaneously making it difficult for an adversary to predict sensitive attributes from the model's outputs, adversarial debiasing can help reduce discriminatory outcomes [33].

In its typical formulation, adversarial debiasing involves two models: a primary predictor and an adversary. The predictor is trained to perform a target task (e.g., predicting recidivism or disease), while the adversary attempts to infer a protected attribute (such as race or gender) from the output of the predictor. Training proceeds in a minimax fashion: the predictor aims to optimize task accuracy while simultaneously minimizing the adversary's ability to detect sensitive attributes. This adversarial game encourages the predictor to become

blind to the protected variable, thereby reducing discriminatory patterns in its decisions. This framework supports multiple definitions of fairness such as demographic parity, equality of odds, and equality of opportunity, and is applicable to both classification and regression tasks [32, pp. 2-5].

2.2 Fair and Unbiased Applications in Healthcare

Recent work in healthcare AI has demonstrated the effectiveness of adversarial debiasing in creating more equitable diagnostic systems. Yang et al. [32] developed an adversarial training framework that significantly reduced bias in clinical machine learning models while maintaining predictive performance. Their approach showed particular promise in addressing racial and gender disparities in healthcare AI systems.

Their model was evaluated across four independent UK National Health Service (NHS) hospital datasets. Notably, it achieved high AUROC scores (0.86–0.89) while significantly improving fairness [32, pp. 3-5], measured by equalized odds across both hospital and ethnic group cohorts. Although the overall diagnostic performance remained consistent between the baseline and adversarial models, fairness improved significantly. For instance, the standard deviation in true positive rates (TPR) and false positive rates (FPR) across ethnic groups declined across validation cohorts. In the Bedfordshire Hospitals cohort, the adversarial model achieved a TPR SD of 0.095 compared to 0.096 in the baseline—demonstrating tighter fairness margins without sacrificing predictive strength [32, p. 4]. Yang et al. [32] also applied t-SNE visualizations to identify clustering by hospital in feature representations, highlighting site-specific bias. The adversarial model disrupted this clustering, suggesting it effectively obfuscated location-based patterns that could lead to unfair treatment.

The COVID-19 pandemic highlighted the urgent need for fair and equitable AI systems in healthcare. Studies showed that biased algorithms could exacerbate existing health disparities, making adversarial debiasing techniques particularly relevant. Correa et al. [12] developed a robust two-step adversarial debiasing approach specifically for medical imaging applications, demonstrating significant improvements in fairness across different demographic groups.

Their approach trains a convolutional neural network (CNN) with a bifurcated structure: a predictor and an adversary, the latter aiming to uncover protected variables like race. During training, they implement gradient reversal and penalization to reduce the network's reliance on those sensitive features. What sets their model apart is partial fine-tuning, wherein only a carefully selected subset of convolutional layers—those most correlated with bias—are updated. This ensures that the model remains fair while retaining diagnostic power. In fact, their partial debiasing model improved AUC scores on some tasks (e.g., pneumothorax and no-finding detection) and reduced TPR disparity in high-risk classes such as Edema from 1.33 to 1.11—well within fairness thresholds defined by the "80% rule" [12, p. 5]. For the mammography dataset, where race is known to correlate with tissue density, partial debiasing preserved classifica-

tion performance while reducing disparity for Asian patients from 3.37 to 1.67 on low-density classes.

2.3 Limitations and Ethical Oversight

These advances in adversarial debiasing represent important progress toward the Digital Humanism goal of technology that serves all members of society equitably. However, technical solutions alone are insufficient; they must be accompanied by broader institutional and policy changes [12].

The integration of fairness considerations into machine learning pipelines requires careful attention to the specific context and stakeholder needs. Different fairness metrics may be appropriate for different applications, and there may be trade-offs between different notions of fairness. The Digital Humanism framework emphasizes the importance of inclusive stakeholder engagement in making these decisions [33].

Still, challenges remain. As noted by Zhang et al. [33], fairness definitions such as demographic parity and equalized odds may not always be compatible, and adversarial training may struggle with convergence in complex models. In medical domains, external validity remains a concern, as datasets often underrepresent minority populations. Nevertheless, these techniques represent a foundational step toward building accountable, transparent, and fair algorithmic systems.

3 Evaluating Bias and Fairness in the COMPAS Algorithm

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm represents one of the most widely studied and controversial applications of predictive analytics in criminal justice. Developed by Northpointe (now Equivant), COMPAS is used across the United States to assess the likelihood of recidivism among defendants and inform decisions about bail, sentencing, and parole [6].

The algorithm gained widespread attention following a 2016 ProPublica investigation that revealed significant racial disparities in its predictions [2, 21]. The investigation found that Black defendants were almost twice as likely as white defendants to be incorrectly flagged as high-risk for reoffending, while white defendants were more likely to be incorrectly flagged as low-risk [8, 13, 17].

These findings sparked intense debate about the use of algorithmic tools in criminal justice and highlighted fundamental questions about fairness in machine learning. The COMPAS controversy illustrates the complex challenges involved in defining and measuring fairness in algorithmic systems [17].

3.1 Data Foundations for Fair and Transparent Modeling

Our previous research has examined various approaches to mitigating bias in recidivism prediction models, including the application of fairness-aware machine learning techniques and adversarial debiasing methods [19, 18]. These studies demonstrated that it is possible to significantly reduce racial disparities in prediction outcomes while maintaining reasonable levels of accuracy.

To better understand and evaluate the fairness implications of COMPAS, we reprocessed the publicly available COMPAS dataset. Initially trained on a dataset comprising 11,757 defendants evaluated in Broward County, Florida, between 2013 and 2014 [6], the dataset included demographic variables such as age, gender, and race, as well as criminal history and judicial features. After cleaning and filtering, the working dataset was reduced to approximately 7,000 records, preserving essential features including race, charge severity, and prior counts [19]. In the next subsection, we are presenting the main findings of our project as they were formulated along three experiments using three different machine learning tools [19, 18].

3.2 Modeling Approaches and Fairness Evaluation

The first experiment employed a Random Forest model with 100 estimators to predict actual two-year recidivism rather than COMPAS scores. Using an 80-20 training-validation split, the model achieved an accuracy of 76.99%. To examine model behavior, LIME (Local Interpretable Model-Agnostic Explanations) was applied to interpret feature importance. The top predictive features were largely temporal: time since last custody, most recent arrest date, and days between screening and arrest. While these features were statistically meaningful, their legal relevance was questionable, suggesting that the model may have overfit on procedural artifacts rather than meaningful behavioral cues [19, 18].

In the second experiment, we implemented adversarial debiasing using the AI Fairness 360 (AIF360) toolkit. A neural network with 200 hidden units was trained to predict recidivism while minimizing the model's dependence on protected attributes (race and sex). The debiased model achieved 76.09% accuracy—a marginal drop from the baseline—yet LIME revealed a shift in feature emphasis. More behaviorally meaningful features such as juvenile felony count and jail release date rose in importance, replacing less interpretable temporal markers. This shift suggests that adversarial training encouraged the model to move away from racially correlated proxy variables.

In a third step, we evaluated both models using fairness metrics from AIF360. The baseline Random Forest model had a demographic parity difference of 0.149 and an equalized odds difference of 0.106—indicating moderate fairness violations. These disparities were reduced under adversarial debiasing, reinforcing the technique's value in improving group-level parity.

3.3 Insights and Human-Centered Implications

The COMPAS case study reveals several important lessons for the development of fair algorithmic systems. First, it demonstrates the importance of comprehensive bias testing across different demographic groups. Second, it highlights the need for transparency in algorithmic decision-making, particularly in high-stakes domains like criminal justice. Third, it underscores the importance of ongoing monitoring and evaluation of deployed systems.

Notably, although neither model explicitly relied on race as a predictor, the baseline model's emphasis on time-related features appears to encode latent correlations with race or socioeconomic status. The disappearance of these artifacts under adversarial debiasing suggests that fairness-aware models can filter out such proxies, contributing to ethically grounded decision-making.

From a Digital Humanism perspective, the COMPAS controversy represents a failure to adequately consider the human and social implications of algorithmic decision-making. The focus on technical accuracy without sufficient attention to fairness and equity led to a system that perpetuated and amplified existing racial disparities in the criminal justice system.

Incorporating interpretability tools such as LIME, fairness evaluation frameworks like AIF360, and adversarial training into model development pipelines is essential. These techniques help ensure that predictive models serve human-centered goals and reflect legal and ethical standards.

Moving forward, the development of recidivism prediction tools must be guided by principles of fairness, transparency, and accountability. This requires not only technical improvements but also broader reforms to ensure that algorithmic tools serve the goals of justice rather than perpetuating discrimination.

4 Interpretability and Explainability in Algorithmic Forensics

The field of forensic science has long grappled with questions of reliability, validity, and interpretability. Traditional forensic methods, from fingerprint analysis to DNA profiling, have faced scrutiny regarding their scientific foundations and the potential for human error and bias. The introduction of algorithmic tools and artificial intelligence into forensic practice brings both opportunities and new challenges for ensuring interpretability and explainability [16].

Interpretability in forensic science refers to the ability to understand and explain how evidence is analyzed and how conclusions are reached. This is crucial not only for scientific validity but also for legal proceedings, where expert testimony must be comprehensible to judges and juries [16]. The challenge becomes more complex when algorithmic tools are involved, as these systems may operate in ways that are difficult for human experts to understand or explain [26, 4].

The 2009 National Research Council report "Strengthening Forensic Science in the United States" highlighted significant concerns about the scientific foundations of many forensic disciplines and called for greater rigor in forensic methods

[22]. These concerns extend to the use of algorithmic tools, which must be subject to the same standards of scientific validity and interpretability.

4.1 Communicating Uncertainty and Misinterpretation of Probabilities

One of the key challenges in forensic science is the communication of uncertainty and the potential for error. Traditional forensic testimony has often presented conclusions with inappropriate certainty, failing to adequately convey the limitations and potential for error in forensic analyses [26]. This problem is compounded when algorithmic tools are used, as these systems may produce probabilistic outputs that are difficult to interpret and communicate [26].

Transparent forensic methods are essential to prevent misinterpretation and maintain trust in legal proceedings. Judges and jurors often lack the technical background required to understand the probabilistic or algorithmic underpinnings of forensic conclusions. Without clear explanations, they may misjudge the weight of evidence, potentially resulting in miscarriages of justice [16].

The prosecutor's fallacy represents a particularly important concern in the interpretation of forensic evidence. This fallacy occurs when the probability of the evidence given innocence is confused with the probability of innocence given the evidence. For example, if a DNA match has a random match probability of 1 in a million, it would be fallacious to conclude that there is only a 1 in a million chance that the defendant is innocent [11].

According to Thompson et al. [26], forensic experts must clearly differentiate between likelihood ratios and posterior probabilities. While a likelihood ratio quantifies how much more likely evidence is under one hypothesis than another, it does not convey the actual probability that a suspect is guilty. Drawing conclusions about guilt from likelihood ratios alone can lead to overconfidence in the evidence and a failure to consider contextual information such as alibis or alternative suspects.

4.2 Fingerprints, Algorithms, and Cognitive Bias

Fingerprint analysis provides an illustrative example of the challenges involved in ensuring interpretability and reliability in forensic science. Traditional fingerprint analysis relies on the subjective judgment of human examiners, who compare latent prints found at crime scenes with known prints from suspects [3]. This process is subject to various sources of error and bias, including confirmation bias, where examiners may be influenced by contextual information about the case [20].

Fingerprint evidence, once considered infallible, has also come under scrutiny. The 2009 National Research Council report and numerous academic studies have questioned the objectivity and reproducibility of traditional fingerprint analysis. Cole (2005) highlighted that fingerprint identification often lacks a meaningful error rate estimate, making it difficult to quantify confidence in a match [26].

The introduction of Automated Fingerprint Identification Systems (AFIS) has brought both benefits and new challenges to fingerprint analysis. These systems can rapidly search large databases and identify potential matches, but they also introduce new sources of potential error and bias [1, 26]. The interpretability of AFIS results is crucial for ensuring that these systems are used appropriately and that their limitations are understood. AFIS also provides match scores and similarity metrics, which enhance interpretability and allow for independent auditing.

Cognitive bias further complicates forensic interpretation. Confirmation bias, for instance, leads analysts to seek out evidence that supports their pre-existing hypotheses while downplaying contradictory data. This is particularly problematic in environments where forensic examiners are exposed to investigative details or expectations from law enforcement [1, 26] documented several real-world cases in which these biases undermined the reliability of forensic testimony.They recommended safeguards such as blind verification, peer review, and the preregistration of analytic methods to minimize subjective influence.

4.3 Challenge of Interpretability

The challenge of interpretability extends beyond individual forensic techniques to the broader question of how forensic evidence is integrated and evaluated. Bayesian approaches to evidence evaluation offer a principled framework for combining multiple sources of evidence and quantifying uncertainty [26]. However, these approaches require careful attention to the assumptions and limitations of the underlying models [26].

DNA analysis represents one of the most scientifically robust forensic techniques, but even here, questions of interpretability and communication remain important. Complex DNA mixtures, degraded samples, and low-level DNA can present significant challenges for interpretation [7]. The use of probabilistic genotyping software has improved the analysis of complex DNA evidence, but these tools require careful validation and interpretation [1].

4.4 Toward a Human-Centered Forensic Framework

The development of interpretable and explainable forensic tools requires a multidisciplinary approach that brings together expertise from forensic science, statistics, computer science, and law [24].

Methodological transparency is also vital: forensic methods must be fully documented, including statistical assumptions, data preprocessing steps, and validation protocols. Agencies like the Scientific Working Group on DNA Analysis Methods (SWGDAM) have provided detailed guidelines for transparent and reproducible forensic practices [7].] Visual aids, analogies, and layperson-friendly language can significantly improve jurors' understanding of complex concepts. For example, using a lottery analogy to explain likelihood ratios can help communicate the rarity of a DNA match in relatable terms [1]. Interpretable models—such as decision trees or rule-based systems—can replace opaque black-

box algorithms in many forensic contexts. Cythia Rudin [24] argues that interpretable models are not only feasible but preferable in high-stakes applications like criminal justice, where understanding and trust are paramount. Bias mitigation strategies, including blind analysis, training in cognitive psychology, and mandatory peer review, should be institutionalized rather than left to individual discretion.

This alignment with the Digital Humanism emphasis on interdisciplinary collaboration and the need to consider the human and social implications of technological tools (National Research Council [22]).

5 Conclusion

The challenges explored in this work—from algorithmic bias in criminal justice to interpretability in forensic science—reflect broader questions about the role of technology in society and the need for approaches that prioritize human welfare and dignity. The Digital Humanism framework provides a valuable lens for understanding these challenges and developing solutions that are both technically sound and ethically grounded.

The expansion of algorithmic systems into domains with high human stakes necessitates a reevaluation of how we define and implement fairness, transparency, and accountability in automated decision-making.

The COMPAS case study demonstrates the real-world consequences of algorithmic bias and the importance of fairness considerations in the development of predictive systems. Experimental findings from adversarial debiasing confirm that these harms can be mitigated without incurring major accuracy losses when models are explicitly trained to disregard protected attributes such as race or gender. In the medical domain, similarly promising results demonstrate that fairness-aware techniques can promote equity in clinical diagnostics, thus reinforcing the cross-domain applicability of these strategies.

The techniques explored for bias mitigation, including adversarial debiasing, offer promising approaches for creating more equitable systems, but they must be accompanied by broader institutional and policy changes.

The challenges of interpretability and explainability in forensic science highlight the importance of transparency and accountability in expert systems. The development of algorithmic tools for forensic applications must be guided by principles of scientific rigor and the need for clear communication of uncertainty and its limitations.

However, fairness is not merely a function of technical optimization; it also demands interpretability and transparency. In forensic science, the absence of clear communication, quantifiable error rates, and protections against cognitive bias has undermined the credibility of expert testimony and, at times, led to wrongful convictions. The adoption of tools like AFIS and the promotion of transparent, interpretable models represent essential progress but much work remains to institutionalize such practices.

Ultimately, achieving algorithmic justice requires more than piecemeal technical fixes. It calls for a systemic commitment to ethical design, legal oversight, and inclusive stakeholder engagement. By embedding fairness, transparency, and accountability into both the design and governance of predictive systems, we can move closer to realizing the full promise of algorithmic technologies while safeguarding the rights and dignity of those they affect.

References

[1] Colin GG Aitken and Franco Taroni. *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley & Sons, 2004.

[2] Julia Angwin et al. "Machine bias". In: *ProPublica* (2016). URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] David R Ashbaugh. *Quantitative-Qualitative Friction Ridge Analysis: An Introduction to Basic and Advanced Ridgeology*. CRC Press, 1999.

[4] David J Balding and Peter Donnelly. "The Prosecutor's Fallacy and DNA Evidence". In: *Criminal Law Review* (1994), pp. 711–721.

[5] Tim Berners-Lee. *30 years on, what's next #ForTheWeb?* 2019. URL: https://onezero.medium.com/30-years-on-whats-next-fortheweb-6ce844ed147f.

[6] Thomas Blomberg et al. *Validation of the COMPAS risk assessment classification instrument*. Tech. rep. Florida State University, 2010. URL: https://criminology.fsu.edu/sites/g/files/upcbnu3076/files/2021-03/Validation-of-the-COMPAS-Risk-Assessment-Classification-Instrument.pdf.

[7] John M Butler. *Advanced Topics in Forensic DNA Typing: Methodology*. Academic Press, 2015.

[8] Alexandra Chouldechova. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: *Big Data* 5.2 (2017), pp. 153–163. DOI: 10.1089/big.2016.0047.

[9] Mark Coeckelbergh. *Introduction to Philosophy of Technology*. Oxford University Press, 2019.

[10] Mark Coeckelbergh. "What is digital humanism? A conceptual analysis and an argument for a more critical and political digital (post)humanism". In: *Journal of Responsible Technology* 17 (2024), p. 100073.

[11] Simon A Cole. "More than Zero: Accounting for Error in Latent Fingerprint Identification". In: *Journal of Criminal Law and Criminology* 95.3 (2005), pp. 985–1078.

[12] Ricardo Correa et al. "A robust two-step adversarial debiasing with partial learning - medical image case-studies". In: *Medical Imaging 2023: Imaging Informatics for Healthcare, Research, and Applications*. Vol. 12469. SPIE. 2023, p. 1246908. DOI: 10.1117/12.2647285.

[13] Julia Dressel and Hany Farid. "The accuracy, fairness, and limits of predicting recidivism". In: *Science Advances* 4.1 (2018), eaao5580. DOI: 10.1126/sciadv.aao5580.

[14] Batya Friedman, Peter H Kahn Jr, and Alan Borning. "Value sensitive design and information systems". In: *Human-computer interaction and management information systems: Foundations*. Ed. by Ping Zhang and Dennis Galletta. New York, NY: Routledge, 2017, pp. 348–372.

[15] Christian Fuchs. *Digital Humanism: A Philosophy for 21st Century Digital Society*. Emerald Publishing, 2022.

[16] Brandon L Garrett and Cynthia Rudin. "Interpretable and Explainable Forensic Science". In: *Annual Review of Criminology* 3 (2020), pp. 1–22.

[17] Ellora Thadaney Israni. *When an Algorithm Helps Send You to Prison*. 2017. URL: https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html.

[18] Rahulrajan Karthikeyan, Chieh Yi, and Moses Boudourides. "Criminal justice in the age of AI: Addressing bias in predictive algorithms used by courts". In: *The Ethics Gap in the Engineering of the Future*. Ed. by Stylios Stelios and Konstantinos Theologou. Leeds: Emerald Publishing Limited, 2024, pp. 27–50. DOI: 10.1108/978-1-83797-635-520241003.

[19] Rahulrajan Karthikeyan et al. "Mitigating bias in judicial ML models to promote fairness in criminal justice system". Unpublished manuscript of CSE 575 course project, Arizona State University. 2023.

[20] Saul M Kassin, Itiel E Dror, and Jeff Kukucka. "The Forensic Confirmation Bias: Problems, Perspectives, and Proposed Solutions". In: *Journal of Applied Research in Memory and Cognition* 2.1 (2013), pp. 42–52.

[21] Jeff Larson et al. "How we analyzed the COMPAS recidivism algorithm". In: *ProPublica* (2016). URL: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[22] National Research Council. *Strengthening Forensic Science in the United States: A Path Forward*. National Academies Press, 2009.

[23] Frank Pasquale. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, 2015.

[24] Cynthia Rudin. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead". In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215.

[25] Jack Stilgoe, Richard Owen, and Phil Macnaghten. "Developing a framework for responsible innovation". In: *Research Policy* 42.9 (2013), pp. 1568–1580.

[26] William C Thompson. "The Role of Probability in Forensic Science". In: *Wiley Encyclopedia of Forensic Science*. Ed. by Allan Jamieson and Andre Moenssens. John Wiley & Sons, 2013.

[27] Vienna Manifesto. *Vienna Manifesto on Digital Humanism*. 2019. URL: https://caiml.org/dighum/dighum-manifesto/.

[28] Hannes Werthner et al. *Introduction to Digital Humanism*. Springer, 2022. URL: https://link.springer.com/book/10.1007/978-3-031-45304-5.

[29] Hannes Werthner et al. "Digital Humanism: The Time Is Now". In: *Computer* 56.1 (2023), pp. 138–142. DOI: 10.1109/MC.2022.3219528.

[30] Langdon Winner. "Do artifacts have politics?" In: *Daedalus* 109.1 (1980), pp. 121–136.

[31] Langdon Winner. *The Whale and the Reactor: A Search for Limits in an Age of High Technology*. University of Chicago Press, 1986.

[32] Jiandong Yang et al. "An adversarial training framework for mitigating algorithmic biases in clinical machine learning". In: *npj Digital Medicine* 6.1 (2023), p. 55. DOI: 10.1038/s41746-023-00805-y.

[33] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. "Mitigating unwanted biases with adversarial learning". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 335–340.