OXFORD UNIVERSITY PRESS | Digital Scholarship in the Humanities

Algorithmic Bias and Fairness: A Digital Humanities Approach to Critical Algorithm Studies

Journal:	*Digital Scholarship in the Humanities*
Manuscript ID	Draft
Manuscript Type:	Short Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Boudourides, Moses; Northwestern University, Data Science Program Karthikeyan, Rahulrajan ; Arizona State University, School of Computing and Augmented Intelligence
Keywords:	Algorithmic Fairness, COMPAS, Interpretability, Explainability, Forensic Science, Adversarial Debiasing, Digital Humanities, Critical Algorithm Studies

Algorithmic Bias and Fairness: A Digital Humanities Approach to Critical Algorithm Studies

Moses Boudourides[1,*] and Rahulrajan Karthikeyan[2]

[1]Data Science Program, School of Professional Studies, Northwestern University, USA
[2]School of Computing and Augmented Intelligence, Arizona State University, USA

*Corresponding author: Moses.Boudourides@northwestern.edu

Abstract

Motivation: Drawing on Digital Humanities approaches to critical algorithm studies, this paper examines how predictive models can perpetuate societal inequities across criminal justice, healthcare, and forensic science domains. Using Digital Humanities theoretical frameworks such as operationalization, environmental scanning, and critical examination of computational tools, this research analyzes algorithmic bias through interdisciplinary perspectives that combine technical rigor with humanistic interpretation. The COMPAS recidivism prediction system serves as a central case study, demonstrating how algorithmic bias can be significantly reduced through fairness-aware machine learning approaches informed by Digital Humanities methodologies. This work extends the inquiry into healthcare, showing how adversarial debiasing improves equity in diagnostic algorithms, particularly during the COVID-19 pandemic. Furthermore, adopting Digital Humanities emphasis on transparency and interpretability, this research examines forensic science applications, 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 demonstrate how Digital Humanities perspectives on computational systems can inform the development of technically robust and ethically grounded algorithmic approaches, capable of delivering not just accurate predictions but justice-aligned outcomes.

Keywords: Algorithmic Fairness, COMPAS, Interpretability, Explainability, Forensic Science, Adversarial Debiasing, Digital Humanities, Critical Algorithm Studies

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 (Chouldechova, 2017; Dressel and Farid, 2018; Angwin et al., 2016; Larson et al., 2016).

This work examines these challenges through the lens of Digital Humanities, an interdisciplinary field that applies computational methods to traditional humanities questions while maintaining critical perspectives on technology's social and cultural implications. Digital Humanities, as defined by scholars like Moretti (2013) and Manovich (2020), emphasizes the importance of combining quantitative analysis with qualitative interpretation, and critically examining how digital tools and datasets shape our understanding of human culture and society.

1.1 Digital Humanities Framework

Digital Humanities offers a critical framework for understanding algorithmic systems within their broader social, cultural, and historical contexts. This field emphasizes the importance of "op-

erationalization"—the process of translating theoretical concepts into computational implementations—and recognizes that these translation processes inevitably embed particular worldviews and assumptions (Pichler and Reiter, 2022). Digital Humanities scholars have developed methodologies for critically examining digital collections and computational tools, including techniques for identifying bias and representational gaps in datasets (Beelen et al., 2023).

In the context of algorithmic decision-making, Digital Humanities provides essential tools for understanding how computational systems shape and are shaped by social structures. The field's emphasis on "distant reading" and cultural analytics offers methods for analyzing large-scale patterns while maintaining attention to local contexts and marginalized voices (Moretti, 2013; Underwood, 2019). This approach is particularly valuable for understanding how algorithmic systems may perpetuate historical inequities embedded in training data.

The Digital Humanities framework also emphasizes the importance of transparency and interpretability in computational methods. Scholars in this field have long grappled with questions of how to make computational processes accessible to humanistic inquiry, developing approaches that balance technical rigor with critical interpretation (Ramsay, 2011). This perspective provides valuable insights for developing algorithmic systems that are not only accurate but also interpretable and accountable to the communities they serve.

2 Literature Review

The study of algorithmic bias and fairness is a rapidly evolving field, drawing on insights from computer science, law, sociology, and critical theory.A foundational concept is that bias is not merely a technical glitch but is often a reflection of deeply embedded societal inequities present in the data used to train machine learning models (Barocas and Selbst, 2016; Noble, 2018). Scholars in Critical Algorithm Studies argue that algorithms are not neutral tools but are instead powerful actors that shape social reality, often in ways that reinforce existing power structures (Gillespie, 2014; Seaver, 2017). This perspective, advanced by thinkers like Safiya Noble and Ruha Benjamin, reveals how seemingly objective systems can produce discriminatory outcomes, a phenomenon Benjamin terms the "New Jim Code" (Benjamin, 2019).

Defining and measuring fairness has proven to be a significant challenge, with computer scientists proposing a variety of mathematical definitions that are often mutually exclusive (Dwork et al., 2012; Corbett-Davies and Goel, 2018). These metrics, such as demographic parity and equality of opportunity, highlight the trade-offs inherent in designing fair systems. The work of Buolamwini and Gebru on intersectional accuracy disparities in facial recognition systems (the "Gender Shades" project) provides a powerful empirical demonstration of how models can fail spectacularly for specific demographic subgroups, underscoring the need for intersectional analysis in fairness audits (Buolamwini and Gebru, 2018).

In response to the "black box" problem, where the internal workings of complex models are opaque, the field of eXplainable AI (XAI) has emerged. However, some scholars, notably Cynthia Rudin, argue for a shift away from post-hoc explanations toward the use of inherently interpretable models, especially in high-stakes decisions (Rudin, 2019). This aligns with the Digital Humanities' emphasis on transparency and the ability to critically interrogate computational processes (Ramsay, 2011). The challenge is not just to explain a model's decision but to ensure that the model itself is based on justifiable and understandable logic.

Case studies provide concrete evidence of algorithmic harm. The ProPublica investigation into the COMPAS algorithm remains a landmark study, revealing significant racial bias in recidivism predictions (Angwin et al., 2016). In healthcare, a widely cited study by Obermeyer et al. found that a commercial algorithm used to predict health needs was systematically biased against Black patients, not because it used race as a variable, but because it used healthcare costs as a proxy for health, failing to account for unequal access to care (Obermeyer et al., 2019). These cases illustrate how seemingly neutral design choices can have profound discriminatory

consequences.

3 A Digital Humanities Lens on Algorithmic Case Studies

Digital Humanities provides a unique and powerful lens for analyzing the case studies at the heart of this paper. By bridging computational methods with critical theory, DH offers a framework for moving beyond purely technical assessments of bias to a more holistic understanding of how algorithms function within complex social systems. The core tenets of DH—skepticism toward data, attention to context, and a demand for interpretability—are directly applicable to the challenges posed by the COMPAS and forensic science algorithms.

In the case of COMPAS, a Digital Humanities approach immediately questions the "operationalization" of recidivism risk. DH scholars are trained to ask how abstract concepts like "risk" are translated into measurable data points and what is lost or distorted in that translation (Pichler and Reiter, 2022). The COMPAS algorithm relies on historical arrest data, which is itself a product of biased policing practices. A DH perspective, informed by methodologies like the "environmental scan" (Beelen et al., 2023), would treat this data not as an objective record of criminality but as a cultural artifact reflecting historical power dynamics. This reframing reveals that the algorithm is not predicting future crime so much as it is predicting future arrests, thereby perpetuating a cycle of surveillance and criminalization for already marginalized communities.

Similarly, when examining forensic algorithms, the DH emphasis on interpretability and its critique of "black boxes" is paramount. The field of "algorithmic criticism" (Ramsay, 2011) insists that we must be able to read and critique the logic of the computational processes we use. The opacity of many forensic software tools, combined with the cognitive biases inherent in their human application, creates a dangerous situation where evidence is presented with an unearned aura of scientific certainty. A DH framework demands that these tools be made transparent and their error rates quantifiable, aligning with the recommendations of the National Research Council (National Research Council, 2009). It provides the critical vocabulary to challenge the prosecutor's fallacy not just as a statistical error, but as a failure of interpretation rooted in an uncritical faith in computational outputs.

4 Addressing Algorithmic Bias through Digital Humanities

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 Humanities perspective, addressing algorithmic bias requires understanding both the technical mechanisms of bias and the historical and social contexts that produce biased data. Digital Humanities scholars have developed sophisticated methods for analyzing bias in cultural datasets, including techniques for identifying underrepresented voices and examining how digitization processes can systematically exclude certain perspectives (Beelen et al., 2023). This section explores various approaches to bias mitigation, with particular attention to adversarial debiasing techniques informed by Digital Humanities methodologies.

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 (Zhang et al., 2018).

Recent work in healthcare AI has demonstrated the effectiveness of adversarial debiasing in creating more equitable diagnostic systems. Yang et al. (2023) 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.

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 (Yang et al., 2023). Correa et al. (2023) developed a robust two-step adversarial debiasing approach specifically for medical imaging applications, demonstrating significant improvements in fairness across different demographic groups.

These advances in adversarial debiasing represent important progress toward the Digital Humanities goal of developing computational methods that are both technically rigorous and socially responsible. However, technical solutions alone are insufficient; they must be accompanied by broader institutional and policy changes and critical examination of the social contexts in which these systems operate (Correa et al., 2023).

5 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 (Blomberg et al., 2010).

The algorithm gained widespread attention following a 2016 ProPublica investigation that revealed significant racial disparities in its predictions (Angwin et al., 2016; Larson et al., 2016). 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 (Chouldechova, 2017; Dressel and Farid, 2018; Israni, 2017).

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 (Israni, 2017).

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 (Karthikeyan et al., 2023, 2024). These studies demonstrated that it is possible to significantly reduce racial disparities in prediction outcomes while maintaining reasonable levels of accuracy.

6 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 (Garrett and Rudin, 2020).

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 (Garrett and Rudin, 2020). 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 (Thompson, 2013; Balding and Donnelly, 1994).

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 (National Research Council, 2009). These concerns extend to the use of algorithmic tools, which must be subject to the same standards of scientific validity and interpretability.

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 (Thompson, 2013). This problem is compounded when algorithmic tools are used, as these systems may produce probabilistic outputs that are difficult to interpret and communicate (Thompson, 2013).

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 (Cole, 2005).

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 (Ashbaugh, 1999). 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 (Kassin et al., 2013).

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 (Aitken and Taroni, 2004; Thompson, 2013).

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 (Butler, 2015). The use of probabilistic genotyping software has improved the analysis of complex DNA evidence, but these tools require careful validation and interpretation (Aitken and Taroni, 2004).

7 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 combine computational rigor with critical social analysis. The Digital Humanities framework provides a valuable lens for understanding these challenges and developing solutions that are both technically sound and contextually informed.

The COMPAS case study demonstrates the real-world consequences of algorithmic bias and the importance of fairness considerations in the development of predictive systems. 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.

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 the design and governance of predictive systems, we can move closer to realizing the full potential of algorithmic technologies while safeguarding the rights and dignity of those they impact. This is, indeed, one of the core humanistic goals that the Digital Humanities seek to advance.

References

Aitken, C. and Taroni, F. (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists*. 2nd edition. John Wiley & Sons, Chichester.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. *ProPublica*, May 23, 2016.

Ashbaugh, D. R. (1999). *Quantitative-Qualitative Friction Ridge Analysis: An Introduction to Basic and Advanced Ridgeology*. CRC Press, Boca Raton, FL.

Balding, D. J. and Donnelly, P. (1994). The prosecutorś fallacy and DNA evidence. *Criminal Law Review*, pages 711–721.

Barocas, S. and Selbst, A. D. (2016). Big dataś disparate impact. *California Law Review*, 104:671–732.

Beelen, K., Wevers, M., Huijnen, P., Heuvel, C. v. d., and Verheul, J. (2023). Bias and representativeness in digitized newspaper collections: Introducing the environmental scan. *Digital Scholarship in the Humanities*, 38(1):1–47.

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press, Cambridge.

Blomberg, T. G., Bales, W. D., Mann, K., Piquero, A. R., and Berk, R. A. (2010). Validation of the COMPAS risk assessment classification instrument. Technical report, Florida State University, College of Criminology and Criminal Justice.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of Machine Learning Research*, volume 81, pages 1–15.

Butler, J. M. (2015). *Advanced Topics in Forensic DNA Typing: Methodology*. Academic Press, San Diego.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163.

Cole, S. A. (2005). More than zero: Accounting for error in latent fingerprint identification. *Journal of Criminal Law and Criminology*, 95(3):985–1078.

Corbett-Davies, S. and Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Correa, R., Jeong, J. J., Patel, B., Trivedi, H., Gichoya, J. W., and Banerjee, I. (2023). A robust two-step adversarial debiasing with partial learning - medical image case-studies. In *Medical Imaging 2023: Imaging Informatics for Healthcare, Research, and Applications*, volume 12469, page 1246908. SPIE.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226.

Garrett, B. L. and Rudin, C. (2020). Interpretable and explainable forensic science. *Annual Review of Criminology*, 3:1–22.

Gillespie, T. (2014). The relevance of algorithms. In *Media Technologies: Essays on Communication, Materiality, and Society*, pages 167–193. MIT Press.

Israni, E. T. (2017). When an algorithm helps send you to prison. *The New York Times*, October 26, 2017.

Karthikeyan, R., Yi, C., and Boudourides, M. (2024). Criminal justice in the age of AI: Addressing bias in predictive algorithms used by courts. In Stelios, S. and Theologou, K., editors, *The Ethics Gap in the Engineering of the Future*, pages 27–50. Emerald Publishing Limited, Leeds.

Karthikeyan, R., Moore, G., Yi, C., Naveen, Y., Madhugondu, D., and Chang, Y.-Y. (2023). Mitigating bias in judicial ML models to promote fairness in criminal justice system. Unpublished manuscript of CSE 575 course project, Arizona State University.

Kassin, S. M., Dror, I. E., and Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition*, 2(1):42–52.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. *ProPublica*, May 23, 2016.

Manovich, L. (2020). *Cultural Analytics*. MIT Press, Cambridge, MA.

Moretti, F. (2013). *Distant Reading*. Verso Books, London.

National Research Council (2009). *Strengthening Forensic Science in the United States: A Path Forward*. The National Academies Press, Washington, DC.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, New York.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Pichler, A. and Reiter, N. (2022). From concepts to texts and back: Operationalization as a core activity of digital humanities. *Journal of Cultural Analytics*, 7(4).

Ramsay, S. (2011). *Reading Machines: Toward an Algorithmic Criticism*. University of Illinois Press, Urbana.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, 4(2).

Thompson, W. C. (2013). The role of probability in forensic science. In Jamieson, A. and Moenssens, A., editors, *Wiley Encyclopedia of Forensic Science*. John Wiley & Sons.

Underwood, T. (2019). *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press, Chicago.

Yang, J., Soltan, A. A. S., Eyre, D. W., Yang, Y., and Clifton, D. A. (2023). An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *npj Digital Medicine*, 6(1):55.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.