Bridging Detection and Fairness in Generative AI: A Cross-Domain Evaluation Approach

Aiko Kato
akto2022@mymail.pomona.edu
Pomona College Computer Science Department
Claremont, California, USA

Abstract

Advances in generative AI have enabled the large-scale production of synthetic text, images, and audio that increasingly resemble human-created content. These developments raise urgent questions about authenticity, fairness, and public trust. Research on AI-generated content detection focuses on distinguishing synthetic from human-made output, while fairness research evaluates whether generative models represent people equitably across demographic groups. Although these areas are usually studied separately, both rely on similar foundations such as dataset quality, consistent evaluation metrics, and transparent reporting. As a result, they tend to suffer from the same underlying weaknesses. This paper argues that detection and fairness should be evaluated together rather than in isolation. Through a cross-domain literature review covering research on detection, debiasing, data governance, and ethical considerations published between 2021 and 2025, the study shows that detection reliability and fairness accountability are closely connected. To address this gap, the paper introduces an evaluation approach that jointly examines how accurately a model detects AI-generated content and how fairly it treats different groups. This integrated perspective aims to support future research and policymaking that promote more reliable, inclusive, and socially responsible generative AI systems.

Keywords

Generative AI, Fairness, AI Detection, Deepfake Detection, Machine Learning Ethics, Data Governance

1 Introduction

Generative AI now produces text, images, and audio that are often indistinguishable from human creations. This progress raises serious concerns about authenticity, fairness, and public trust. As AI systems shape journalism, education, and the creative industries, society must confront new questions about what is genuine and who is represented. The central problem is that current methods cannot reliably distinguish AI-generated content from human work or ensure a diverse and inclusive representation. These challenges show that current evaluation methods are not adequate for either authenticity or representation.

Existing research in AI content detection focuses on measurable features such as linguistic perplexity or frequency patterns in visual data. Work on fairness, in contrast, relies on demographic evaluation of generated outputs to assess representation. Each field faces similar obstacles, including limited and biased datasets, inconsistent evaluation metrics, and opaque model architectures that resist interpretation. These weaknesses make it difficult to produce reliable results or to compare findings across studies. Although both areas have advanced independently, their separation obscures how the same data and measurement flaws compromise both technical reliability and ethical soundness. In both fields, these weaknesses arise from flaws in evaluation design rather than model architecture.

These shared limitations suggest that detection and fairness should not be evaluated separately but understood as parts of the same evaluation problem. If detection algorithms misclassify text or images due to biased training data, they reproduce the same inequities that fairness researchers seek to correct. Therefore, improving trust in generative AI requires a combined approach that links detection performance with representational ethics. This connection forms the starting point of this study, which treats technical accuracy and ethical accountability as mutually reinforcing goals.

This project will synthesize literature from major AI conferences and journals published between 2021 and 2025. The analysis will review methods for detecting AI-generated text, images, and audio; fairness frameworks designed to evaluate diversity and mitigate bias; data governance practices that ensure reliability; and societal and legal contexts that shape ethical AI development. It will use a cross-domain mapping approach that systematically compares how the same structural weaknesses (dataset bias, classifier error, distribution shift) appear across detection and fairness pipelines and examines how dataset quality, evaluation consistency, and model transparency influence both detection reliability and fairness outcomes. The result will be a comparative analysis that builds a conceptual bridge between technical and ethical evaluation criteria.

The anticipated outcome is an integrated evaluation approach that enables simultaneous assessment of detection accuracy and fairness accountability in generative AI systems. This approach will provide researchers with concrete criteria for identifying when detection tools disproportionately flag content from marginalized groups or when debiasing methods inadvertently compromise detection robustness. Key challenges include the rapid evolution of AI models, which may quickly outpace detection methods, and cultural variation in fairness expectations. By focusing on transparent evaluation criteria, this work aims to support generative AI systems that are more reliable, more inclusive, and more deserving of public trust.

Figure 1 provides a simplified overview of the parallel detection and fairness pipelines introduced in this section. Although the two pipelines appear separate, the figure shows that they depend on the same components such as dataset quality, auxiliary classifiers, and evaluation metrics, which means failures in these components can propagate across both tasks. The unified evaluation framework addresses this by requiring joint evaluation, group-disaggregated metrics, and fairness estimates that account for classifier error.

This structure makes it possible to identify tradeoffs and hidden interactions that isolated evaluation pipelines cannot detect.

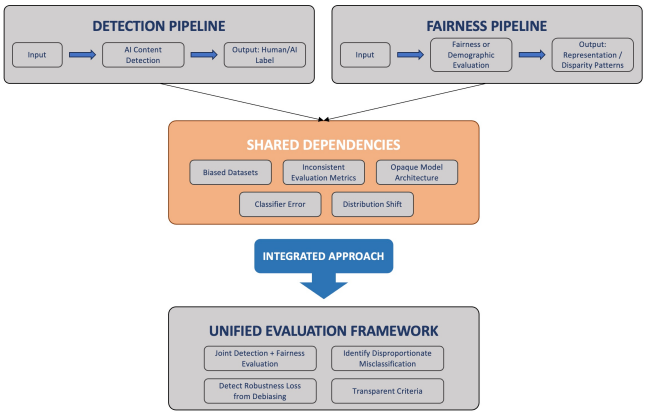

Figure 1: Overview of the detection and fairness pipelines, their shared structural weaknesses, and the proposed unified evaluation framework.

Prior research evaluates detection and fairness separately, which hides how errors in one domain distort results in the other. No existing work evaluates how debiasing affects detectability, how classifier error propagates into both detection and fairness metrics, or how group-specific misclassification rates change when the same dataset is used for both tasks. The contribution of this paper is to formalize these cross-domain dependencies and introduce a unified evaluation framework that analyzes detection accuracy and fairness on the same dataset, reports their results together, and incorporates basic controls for classifier error and uncertainty. By making these dependencies visible, the framework reveals tradeoffs and measurement failures that isolated evaluation cannot detect.

2 Literature Review

This section reviews three bodies of work that parallel the goals of this project: (1) AI content detection methods, (2) bias and fairness in generative models, and (3) reliability and data governance. Across these strands, a common pattern emerges: technical proposals for detection or fairness are limited less by model architecture than by dataset quality, evaluation design, and the absence of cross-domain approaches that connect detection reliability with representational fairness.

2.1 AI Content Detection Methods

Research on detecting AI-generated content spans text, images, video, and multimodal media. For text, Tulchinskii et al. [15] propose the Persistent Homology Dimension (PHD) detector, which uses tools from topology to measure the intrinsic dimension of text representations. They show that human-written text occupies a higher-dimensional manifold than AI-generated text, enabling a model-independent detector that remains robust across multiple language models and languages and that reduces bias against non-native English writers. This work illustrates how interpretable, feature-level reasoning can improve both accuracy and fairness in text detection. However, while Tulchinskii et al. claim their detector reduces bias against non-native English writers, they provide no demographic breakdown of false positive rates. This is a significant omission given that their entire fairness claim rests on this untested assumption. If human writing complexity varies systematically across demographic groups, the PHD detector's reliance on dimensional complexity could inadvertently encode demographic signals, undermining its claimed fairness advantages.

For visual deepfakes, Yan et al. [17] introduce DeepfakeBench, a standardized benchmark that unifies datasets, training scripts, and evaluation protocols for fifteen detection models across nine datasets. Their experiments reveal that simple CNN baselines can match or exceed more sophisticated models when evaluated fairly, showing that fragmented datasets and inconsistent metrics obscure our understanding of detection performance. Taken together, these results indicate that recent work may have emphasized architectural novelty while overlooking foundational evaluation issues. If simple CNNs can match complex models under fair evaluation, this raises uncomfortable questions about whether recent detection research has delivered genuine progress or merely optimized for fragmented benchmarks. Zegeye et al. [18] extend this perspective with a survey of machine-learning-based deepfake detection, comparing classical ML models with deep networks and emphasizing how accuracy drops sharply when moving from high-quality benchmark data to noisy, real-world media.

Salman et al. [13] provide a complementary overview that categorizes deepfakes (face-swapped, voice-cloned, fully synthetic, etc.) and surveys multimodal detection methods that combine audio and visual signals. While their proposed multimodal detector outperforms earlier systems, it still achieves modest accuracy in realistic settings, reinforcing concerns from Mirsky and Lee [12] that the arms race between deepfake generation and detection leads to rapidly outdated detectors and persistent vulnerabilities. Taken together, these works demonstrate impressive progress in detecting synthetic content, but they also underscore recurring weaknesses: overfitting to specific datasets, lack of robustness to distribution shift, and limited discussion of how detection errors may unevenly impact different demographic groups. The modest 66.5% accuracy of Salman et al.'s multimodal detector indicates that, even with audio and visual signals combined, it barely outperforms simpler methods. This suggests that current approaches may be reaching fundamental limits rather than needing incremental refinement. Most critically, none of these detection studies systematically examine whether their errors are demographically skewed, an oversight that becomes more troubling when we consider the fairness evaluation problems discussed in the next section.

2.2 Bias and Fairness in Generative Models

A parallel body of work studies bias and fairness in generative models. Mehrabi et al. [11] and Caton and Haas [2] provide foundational surveys that map sources of bias (measurement, representation, sampling, historical, and algorithmic) and organize mitigation strategies into pre-processing, in-processing, and post-processing interventions. They also catalog fairness definitions such as demographic parity and equalized odds and emphasize that these metrics can conflict, forcing practitioners to confront tradeoffs

between accuracy, different notions of fairness, and domain-specific norms. While these comprehensive surveys provide essential taxonomies of bias types and mitigation strategies, their emphasis on mathematical fairness definitions may inadvertently obscure deeper normative questions: Who gets to define what counts as fair? In what contexts do different fairness criteria apply? These surveys treat fairness as primarily a technical optimization problem, but as LaRosa and Conklin later argue through contextual integrity, fairness judgments are inseparable from social context and power dynamics.

Building on this foundation, Teo et al. [14] examine how fairness is actually measured in generative models and uncover a critical weakness: most fairness audits rely on auxiliary classifiers to infer demographic attributes from generated images. Even when these classifiers are highly accurate, their residual errors can introduce fairness estimation errors of up to 17 percentage points, calling into question many published fairness claims. Their CLEAM framework corrects for classifier error using statistical modeling, revealing persistent gender and racial bias in models such as StyleGAN2, StyleSwin, and Stable Diffusion. These results imply that fairness evaluation pipelines inherit failures from detection-like components, reinforcing that the two cannot be treated in isolation. Teo et al.'s findings are significant for the fairness literature: if even highly accurate classifiers introduce estimation errors up to 17 percentage points, then many published claims about fairness improvements may be illusory. This calls into question the credibility of many fairness audits, including evaluations of SFID and Fair DiFFusion, which depend on the same unreliable attribute classifiers highlighted in Teo et al.'s study.

Jung et al. [7] propose Selective Feature Imputation for Debiasing (SFID), a unified debiasing method for vision–language models that identifies bias-associated features and replaces them with more neutral representations. SFID reduces gender and racial bias across classification, captioning, and text-to-image tasks without retraining the underlying models. Friedrich et al. [5] introduce Fair DiFFusion, a post-deployment fairness instruction framework for Stable Diffusion that allows users to steer demographic proportions (e.g., gender and age in occupations) at inference time while maintaining image quality. Both studies demonstrate that debiasing is technically feasible, but also implicitly rely on the kinds of fairness metrics that Teo et al. show to be error-prone. This tension motivates a more integrated view of detection, attribute classification, and fairness auditing. Furthermore, neither Jung et al. nor Friedrich et al. address a critical question: do debiased models become harder to detect as synthetic? If bias-associated features also serve as detection signals, such as when gender stereotypes in generated images help distinguish them from photographs, then debiasing could inadvertently compromise detection reliability. This unexplored tradeoff exemplifies why detection and fairness must be evaluated jointly rather than in isolation. These findings indicate that fairness evaluations depend heavily on the stability of auxiliary classifiers, and that small classification errors can cause large swings in measured fairness outcomes.

2.3 Reliability and Data Governance

A growing literature frames trustworthy AI as a problem of reliability and data governance rather than architecture alone. Liang et al. [10] argue that high-quality, well-documented datasets are the foundation of trustworthy AI, proposing tools such as data nutrition labels and data valuation techniques to diagnose bias and improve representativeness. A data nutrition label is a structured summary of a dataset's contents, including demographics, collection methods, known biases, and documented limitations, which helps researchers understand its quality and risks. While Liang et al.'s data-centric perspective is compelling, they provide limited guidance on practical adoption: how would data nutrition labels be enforced? Who would conduct data valuation? More fundamentally, can static documentation tools address the adversarial dynamics that Juefei-Xu et al. [6] describe, where datasets become obsolete as quickly as they are created? These implementation challenges suggest that data governance requires not just better tools but sustained institutional commitment and regulatory frameworks.

Wang et al. [16] analyze deepfake detection from a reliability perspective, introducing statistical methods that use confidence intervals to quantify how stable detector accuracy is across datasets and perturbations. They find that average accuracies around 69% mask large uncertainty under distribution shift, raising doubts about whether current detectors are forensically reliable. This suggests that deepfake detectors are not yet reliable enough for high-stakes applications such as legal evidence, journalistic verification, or content moderation. Wang et al.'s statistical rigor exposes an uncomfortable gap between the confident claims of many detection papers and the actual stability of their results under realistic conditions. Their confidence interval methodology offers a template for the joint reliability-fairness metrics proposed in Section 3, but must be extended to capture demographic disaggregation alongside statistical uncertainty.

Juefei-Xu et al. and Kaur et al. [8] both emphasize the dynamic, adversarial nature of the deepfake ecosystem: as generative models improve, detectors quickly become obsolete, and benchmark datasets lag behind real-world manipulation techniques. Kaur et al. focus in particular on efficiency and real-time detection constraints, showing that models that appear strong in offline experiments often fail under deployment constraints. Across these works, data governance, continuous evaluation, and transparent reporting emerge as prerequisites for any system that claims to be both reliable and fair. The adversarial arms race that Juefei-Xu et al. document reveals an important challenge: detection is not a problem to be solved once but an ongoing process of adaptation. Kaur et al.'s emphasis on deployment constraints further highlights how laboratory performance can be misleading. Models that achieve impressive offline accuracy often fail under real-time or resource-constrained conditions. Together, these works demonstrate that reliability is not merely a technical property but an ecological one, dependent on evolving threats, deployment contexts, and maintenance infrastructures. Taken together, these works show that reliability research lacks consistent evaluation practices, which makes it difficult to compare results or assess progress across studies.

Across detection, fairness, and reliability research, the same structural weaknesses reappear. Dataset imbalance affects both misclassification rates and representation measures. Attribute classifiers used for fairness evaluation introduce instability that also harms detection performance. Distribution shift reduces both detection accuracy and fairness estimates when models are evaluated outside controlled benchmark settings. These parallels show that detection and fairness cannot be reliably assessed in isolation.

Taken together, these three strands of research reveal substantial progress but also recurring limitations. Detection studies expose challenges in robustness and evaluation design, while fairness studies show that demographic audits depend heavily on fragile attribute classifiers. Reliability and governance work identifies deeper structural issues but does not connect them to fairness outcomes. None of these areas explicitly examine how weaknesses in detection pipelines propagate into fairness metrics, nor how fairness interventions may inadvertently alter detection performance. This gap motivates the central contribution of this paper: an integrated evaluation approach that links detection reliability, classifier stability, demographic fairness, and uncertainty reporting. Section 3 outlines how this approach builds on prior work while addressing the structural interdependencies that existing research treats separately.

3 Proposed Research Contribution: An Integrated Evaluation Approach

This paper brings together two research areas, AI content detection and fairness in generative models, which are usually evaluated separately even though they depend on similar assumptions about data quality, classifier stability, and evaluation design. By synthesizing recent findings across these areas, the study identifies structural weaknesses that influence both detection reliability and fairness outcomes. Building on this shared foundation, the paper introduces an integrated evaluation approach that examines detection performance and representational fairness through group-disaggregated metrics, sensitivity checks for classifier error, and transparent reporting of uncertainty across datasets. This approach clarifies how technical and ethical factors interact and provides a consistent basis for evaluating generative AI systems in ways that support more reliable and accountable development. At a practical level, the integrated approach evaluates detection accuracy and fairness on the same datasets, reports their metrics together, and interprets them as connected outcomes rather than separate assessments. The workflow follows a simple sequence: run detection and fairness evaluations on the same dataset, record their metrics in the same evaluation table, and interpret their results jointly.

3.1 Research Objectives

The study pursues four main objectives:

(1) Identify shared vulnerabilities in detection and fairness evaluation, focusing on how dataset imbalance, classifier error, and distribution shift weaken both areas.
(2) Show that improvements in one domain, such as debiasing generative models, can unintentionally reduce performance or stability in another, such as detection reliability.
(3) Develop an integrated evaluation approach that combines detection performance, demographic fairness, robustness under domain shift, and transparency of data and models.
(4) Provide practical guidance for researchers on reporting standards, uncertainty disclosure, and evaluation practices that link technical accuracy with fairness accountability.

These objectives move beyond existing literature by treating detection, fairness, and reliability as interdependent components of a single evaluation problem.

3.2 Research Design and Methods

(1) **Comparative Literature Synthesis.** Compare detection methods, fairness and debiasing approaches, and reliability/data-governance frameworks. Focus on how each defines success, what datasets and metrics they rely on, and where fairness and reliability concerns are acknowledged or overlooked.
(2) **Cross-Domain Mapping.** Identify shared failure modes across detection and fairness pipelines, including biased datasets, dependence on auxiliary classifiers, brittleness under distribution shift, and gaps in governance. Highlight how these weaknesses undermine both technical and ethical claims.
(3) **Evaluation Approach Development.** Propose an integrated evaluation approach that specifies reporting standards for datasets and classifier accuracy, integrates detection and fairness metrics, includes robustness checks across datasets and perturbations, and incorporates qualitative considerations such as consent, contextual integrity, and potential harms.
(4) **Metrics and Evaluation Components.** Existing detection and fairness studies rely on a shared pool of metrics that can be integrated into a unified evaluation framework. For detection, common metrics include accuracy, precision, recall, AUROC, and the rates of false positives and false negatives. These metrics often change considerably when models are evaluated under distribution shift. In deepfake detection, researchers also report cross-dataset generalization scores and perturbation robustness. Fairness evaluations typically use demographic parity difference, equalized odds, group-wise FPR/FNR, and proportional representation measures for generated outputs. However, these metrics depend on auxiliary attribute classifiers whose error propagates into fairness estimates. The proposed framework incorporates both families of metrics by requiring: (a) group-disaggregated detection performance (per demographic subgroup), (b) fairness metrics corrected for classifier error (e.g., via CLEAM-like adjustments), and (c) uncertainty estimates such as confidence intervals for detector stability. Introducing these metrics directly into the evaluation pipeline clarifies the interactions between technical accuracy and representational fairness by revealing when high detection accuracy coincides with demographic disparities in error rates or when fairness improvements compromise detection robustness. This framework does not produce a single combined score; instead, it reports multiple

complementary metrics, including detection performance, group-disaggregated error rates, classifier corrected fairness estimates, and uncertainty intervals.

(5) **Validation Strategy.** This integrated evaluation approach will be validated using three criteria. First, it must reveal dependencies between detection performance and fairness outcomes that isolated evaluation fails to detect. For example, the framework should show whether debiasing methods change detection accuracy for specific demographic groups. Second, it must improve measurement reliability by incorporating classifier-error correction and uncertainty estimates, resulting in fairness and detection metrics with clearly reported confidence intervals. Third, it must demonstrate practical applicability. The framework should be implementable with existing datasets and publicly available tools, with clear documentation that other researchers can follow. Validation will involve applying the framework to at least two published generative models and showing that it identifies tradeoffs or measurement artifacts that were not visible in the models' original evaluations.

To illustrate how this framework would operate in practice, consider the following hypothetical scenario. Suppose a detector reports 92% overall accuracy. When evaluated with group-disaggregated metrics, the system shows a 14-point higher false-positive rate for older adults than for young adults. A fairness audit of the same model, using an attribute classifier with a 6% error rate, reports that demographic parity improved after debiasing. However, CLEAM-corrected estimates show that the improvement was an artifact of classifier error. Joint evaluation on the same dataset reveals a hidden tradeoff: the debiasing method reduced demographic disparity in synthetic image outputs but also made the model 11% easier to fool for certain subgroups. These interactions would be invisible under isolated evaluation pipelines, which illustrates why integrated assessment is necessary. This proposal will investigate whether such patterns actually occur when the integrated framework is applied to existing generative models. The goal is not to solve detection or fairness, but to expose their dependencies by reframing detection reliability, fairness measurement, and data governance as a single coupled evaluation problem for generative AI.

4 Ethical Implications

Technical work on detection and fairness operates within broader societal and legal concerns. Research in algorithmic governance shows that accountability and trust cannot be resolved through technical accuracy alone, as emphasized by Das et al. [4]. LaRosa and Conklin [9] argue that synthetic media raises questions about contextual integrity, privacy, and the protection of personal likeness, highlighting that ethical evaluation depends on how AI-generated content interacts with existing social norms. Deepfake studies further show that these systems exploit social rather than purely technical vulnerabilities, contributing to harassment, manipulation, and erosion of trust in digital evidence, as noted by Mirsky and Lee [12]. Chesney and Citron [3] further demonstrate that deepfakes facilitate political manipulation, enable targeted harassment, and disproportionately threaten marginalized individuals. These perspectives show that evaluation of detection and fairness must consider the societal contexts in which generative AI is deployed.

These concerns manifest clearly when evaluation methods are unreliable. For instance, an AI detector used in educational settings could incorrectly flag non-native English writers at higher rates, a plausible outcome given patterns in linguistic complexity discussed by Tulchinskii et al. [15]. This aligns with findings by Buolamwini and Gebru [1], who show that demographic misclassification disproportionately affects darker-skinned individuals and women. Detection tools can also be repurposed for surveillance, while fairness interventions that adjust demographic outputs may conflict with cultural or legal privacy expectations. Power dynamics influence these outcomes, since researchers and platforms define which demographic categories matter and which tradeoffs are acceptable. Although the integrated evaluation approach cannot resolve these normative questions, it can make underlying assumptions visible. By requiring group-disaggregated metrics, uncertainty reporting, and a contextual discussion of deployment risks, the approach supports more accountable and socially aware use of generative AI systems. These ethical risks reinforce the need for the integrated evaluation approach proposed in this paper. By requiring demographic disaggregation, classifier-error correction, and explicit uncertainty reporting, the framework makes it possible to detect when technical measurement failures produce social harms, particularly for groups that are already vulnerable to misclassification, over-surveillance, or representational erasure.

5 Conclusion

Generative AI increasingly blurs the line between human and synthetic content, raising pressing questions about authenticity, representation, and trust. Although detection and fairness research have advanced quickly, evaluating them in isolation hides the fact that they are weakened by the same structural issues.

This study shows that detection reliability and representational fairness are not separate concerns but interdependent. The proposed integrated evaluation approach addresses this connection by requiring disaggregated performance metrics, consistent reporting standards, and clear documentation of measurement uncertainty. These components make it possible to identify when detection systems disproportionately flag certain groups or when fairness interventions affect detection robustness. By bringing these evaluation practices together, the approach supports more reliable and culturally aware development of generative AI systems.

Important challenges remain. Generative models evolve rapidly, fairness expectations differ across cultural contexts, and governance structures must balance innovation with accountability. Addressing these open questions will require collaboration across technical, social, and regulatory domains. By making dependencies and tradeoffs visible rather than implicit, this work offers a foundation for more trustworthy and socially responsible generative AI.

6 Acknowledgements

I used Grammarly and LaTeX grammar checkers to correct typos and grammatical errors, and I used ChatGPT to help rephrase awkward or unclear phrasing.

References

[1] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*).* 77–91.

[2] Simon Caton and Christian Haas. 2024. Fairness in Machine Learning: A Survey. *Comput. Surveys* 56, 7, Article 166 (2024).

[3] Robert Chesney and Danielle K. Citron. 2019. Deepfakes and the New Disinformation War. *Foreign Affairs* (2019).

[4] Sanmay Das et al. (Eds.). 2024. *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES-24).* Vol. 7. AAAI Press.

[5] Felix Friedrich et al. 2024. Auditing and Instructing Text-to-Image Generation Models on Fairness. *AI and Ethics* 5 (2024), 2103–2123.

[6] Felix Juefei-Xu et al. 2022. Countering Malicious DeepFakes: Survey, Battleground, and Horizon. *International Journal of Computer Vision* 130 (2022), 1678–1734.

[7] Hoin Jung et al. 2024. A Unified Debiasing Approach for Vision-Language Models across Modalities and Tasks. In *Advances in Neural Information Processing Systems*, Vol. 38.

[8] Achhardeep Kaur et al. 2024. Deepfake Video Detection: Challenges and Opportunities. *Artificial Intelligence Review* 57, Article 159 (2024).

[9] Emily LaRosa and Sherri Conklin. 2025. Deepfakes and Contextual Integrity Violations. In *Proceedings of the 2025 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS).* IEEE.

[10] Weixin Liang et al. 2022. Advances, Challenges and Opportunities in Creating Data for Trustworthy AI. *Nature Machine Intelligence* 4 (2022), 669–677.

[11] Ninareh Mehrabi et al. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6, Article 115 (2021).

[12] Yisroel Mirsky and Wenke Lee. 2020. The Creation and Detection of Deepfakes: A Survey. *Comput. Surveys* 54, 1, Article 7 (2020).

[13] Sonia Salman et al. 2023. Deep Fake Generation and Detection: Issues, Challenges, and Solutions. *IT Professional* 25, 1 (2023), 52–59.

[14] Christopher T. H. Teo et al. 2023. On Measuring Fairness in Generative Models. In *Advances in Neural Information Processing Systems*, Vol. 37.

[15] Eduard Tulchinskii et al. 2023. Intrinsic Dimension Estimation for Robust Detection of AI-Generated Text. In *Advances in Neural Information Processing Systems*, Vol. 37.

[16] Tianyi Wang et al. 2024. Deepfake Detection: A Comprehensive Survey from the Reliability Perspective. *Comput. Surveys* 57, 3, Article 58 (2024).

[17] Zhiyuan Yan et al. 2023. DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection. In *Advances in Neural Information Processing Systems*, Vol. 37.

[18] Bethel Zegeye et al. 2025. Deepfake Detection Using Machine Learning: A Comprehensive Literature Review. In *Proceedings of the 2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA).* IEEE.