
Ethical Implications of AI-Powered Courtroom Transcription Systems

Dr. Adrian Cole Marwick

Associate Professor, Department of Law,
Artificial Intelligence and Digital Justice, Northbridge Metropolitan University,
Seattle, Washington, United States

Corresponding Author: Dr. Adrian Cole Marwick

Associate Professor, Department of Law,
Artificial Intelligence and Digital Justice, Northbridge Metropolitan University,
Seattle, Washington, United States

Article Type: Simulation-Based Research Article

Submitted: 03- December-2024 **Revised:** 10- December-2024

Accepted: 17- December-2024 **Published:** 24- December-2024

Abstract: Artificial intelligence-powered courtroom transcription systems are transforming the production, retrieval, and management of judicial records by converting spoken proceedings into searchable digital text. Although automatic speech recognition offers substantial potential for improving transcription speed, accessibility, and administrative efficiency, its adoption within judicial environments raises significant ethical concerns because transcription errors can influence evidentiary interpretation, appellate review, litigant rights, and public trust. This study examines the ethical implications of AI-assisted courtroom transcription through a simulation-based analytical framework emphasizing accuracy, linguistic fairness, procedural justice, human oversight, privacy, accountability, and contestability. A synthetic audit of 800 courtroom-style transcription instances was designed across four speech conditions: Mainstream US English, African American English, South Asian-accented English, and disfluent speech. Word Error Rate and critical legal transcription error were used as principal performance indicators. The simulated results demonstrated substantial variation in error patterns, with higher modeled error rates under linguistically and acoustically complex conditions.

Logistic regression indicated that increasing Word Error Rate and overlapping speech were associated with elevated probabilities of legally consequential errors, whereas structured human verification substantially reduced modeled risk. The paper proposes a Human-

Accountable Courtroom Transcription Governance Model in which artificial intelligence produces provisional rather than automatically authoritative records. Responsible implementation requires subgroup performance audits, secure audio governance, uncertainty disclosure, traceable corrections, professional human verification, meaningful appeal mechanisms, and periodic institutional auditing.

Keywords: artificial intelligence; courtroom transcription; automatic speech recognition; algorithmic fairness; procedural justice; legal technology; human oversight.

Introduction

Artificial intelligence has become increasingly influential within judicial administration, supporting document classification, legal research, scheduling, evidence management, translation, and transcription. Automatic Speech Recognition (ASR) systems are particularly attractive because they can transform lengthy courtroom proceedings into searchable textual records with significantly reduced processing time.

Courtroom transcription, however, cannot be treated as an ordinary speech-to-text application. Judicial transcripts may influence appellate proceedings, evidentiary interpretation, attorney preparation, accessibility accommodations, and determinations concerning what was actually stated during litigation. A transcription error involving a negation, numerical value, personal name, statutory provision, admission, or speaker identity may therefore have consequences substantially greater than an ordinary linguistic mistake.

Major ethical concerns include:

- unequal transcription accuracy across accents, dialects, and speaking patterns
- incorrect attribution of statements to courtroom participants
- hallucinated or inserted words not present in the source audio
- inappropriate exposure of confidential courtroom information
- excessive reliance on apparently authoritative machine-generated text
- uncertainty regarding responsibility when automated transcripts contain errors
- insufficient mechanisms for challenging inaccurate records

Accordingly, the central policy question is not simply whether AI can transcribe legal proceedings efficiently. The more important question is whether AI-generated transcripts can

be incorporated into judicial records without undermining fairness, accountability, transparency, and due process.

Literature Review

a) Accuracy, Dialect, and Algorithmic Fairness

ASR performance is not uniformly distributed across speakers. Koenecke et al. demonstrated substantial racial disparities in commercial speech-recognition systems, while Tatman and Markl identified performance differences associated with dialect, gender, and regional linguistic variation. Such findings are particularly important for judicial environments because differential accuracy can translate technical inequality into procedural inequality.

Important concerns identified within the literature include:

- accent-sensitive recognition errors
- reduced performance for nonstandard linguistic varieties
- demographic disparities hidden by aggregate accuracy measures
- inaccurate recognition of specialized terminology
- limited transparency concerning subgroup-level performance

b) Acoustic Complexity and Legally Material Errors

Courtrooms commonly involve rapid exchanges, interruptions, background noise, overlapping speakers, legal terminology, names, numerical evidence, and emotionally distressed speech. Loakes demonstrated that indistinct forensic audio can substantially undermine automatic transcription reliability.

Koenecke and colleagues further documented speech-to-text hallucinations in which transcription systems generated phrases absent from the original audio. In judicial settings, such failures may alter the interpretation of testimony or introduce statements never spoken.

Research Gap

Existing scholarship provides substantial evidence that automatic speech-recognition systems can exhibit uneven accuracy across linguistic and demographic groups. Research has also

demonstrated that acoustic degradation, overlapping speech, accent diversity, dissiliency, and specialized vocabulary can decrease transcription performance.

Nevertheless, several gaps remain.

First, much existing research examines technical accuracy without sufficiently connecting transcription errors with procedural justice and evidentiary integrity.

Second, Word Error Rate remains the dominant performance measure despite its inability to distinguish between trivial and legally consequential mistakes. Incorrectly transcribing an article or punctuation mark is fundamentally different from omitting the word “not,” changing a monetary amount, or assigning testimony to the wrong speaker.

Third, limited scholarship establishes operational boundaries between AI-generated preliminary text and legally authoritative judicial records.

Fourth, technical assessments often focus on average accuracy and therefore risk concealing subgroup disparities.

The present study addresses these gaps by introducing the concept of Critical Legal Transcription Error, defined as an error capable of materially modifying the legal interpretation, evidentiary meaning, speaker attribution, or procedural significance of recorded courtroom speech.

The study consequently combines technical accuracy, error materiality, subgroup fairness, human oversight, transparency, auditability, and contestability within a unified ethical framework.

Novelty / Proposed Contribution

The principal contribution of the study is the proposed Human-Accountable Courtroom Transcription Governance Model (**HACT-GM**).

The framework treats automated transcription as an assistive technological layer rather than an autonomous authority responsible for creating the definitive judicial record.

Five dimensions distinguish the proposed model from conventional ASR evaluation:

- **Technical Fidelity:** Measurement of how accurately spoken language is represented.
- **Legal Materiality:** Assessment of whether errors alter legally meaningful content.

- **Distributional Fairness:** Evaluation of whether particular speech communities experience disproportionate error rates.
- **Contestability:** Provision of mechanisms allowing authorized participants to challenge or correct disputed transcript passages.
- **Institutional Accountability:** Preservation of identifiable human and organizational responsibility throughout the transcription workflow.

The framework additionally recommends that courts distinguish among three transcript statuses:

- **Machine Draft** → automatically produced and clearly marked as unverified.
- **Human-Verified Transcript** → compared with source audio and corrected by an authorized professional.
- **Official Judicial Record** → formally approved under applicable institutional procedures.

This tiered structure reduces the risk that visually polished AI output will be mistakenly treated as an infallible representation of courtroom speech.

Research Objectives / Research Questions

The principal objective of this research is to evaluate the ethical risks associated with AI-powered courtroom transcription and determine whether structured human oversight can reduce legally consequential transcription failures.

Specific objectives are:

1. To evaluate simulated transcription accuracy under heterogeneous speech conditions.
2. To examine disparities in Word Error Rate across different modeled speech groups.
3. To distinguish ordinary transcription mistakes from critical legal transcription errors.
4. To evaluate whether overlapping speech increases the probability of critical errors.
5. To determine the protective effect of structured human verification.
6. To develop an accountable governance framework for judicial ASR deployment.

The research questions are:

RQ1: How does simulated Word Error Rate vary across heterogeneous courtroom speech conditions?

RQ2: Does increasing Word Error Rate increase the probability of critical legal transcription errors?

RQ3: Does overlapping speech significantly increase transcription risk?

RQ4: Can structured human verification reduce legally consequential transcription errors?

RQ5: What governance mechanisms should regulate AI-generated courtroom transcripts before they acquire official status?

Methodology

a) Research Design and Simulated Dataset

The research employed a simulation-based quantitative ethical audit. A synthetic dataset containing 800 courtroom-style transcription instances was constructed and divided equally among four speech-condition categories. The statistical values are illustrative and do not represent measurements from actual litigants or commercial ASR platforms.

Table 1: Research Variables and Operationalization

Variable	Operational Definition	Measurement	Analytical Role
Transcript Instances	Synthetic courtroom speech segments	n = 800	Unit of analysis
Speech Condition	Four modeled linguistic conditions	Categorical	Fairness predictor
Word Error Rate	Insertions + deletions + substitutions relative to reference words	Percentage	Primary accuracy metric
Critical Legal Error	Error materially affecting legal meaning	Binary	Primary outcome
Overlapping Speech	Simultaneous speaker interference	Yes/No	Risk predictor
Background Noise	Simulated acoustic interference	Yes/No	Risk predictor
Human Verification	Professional review of machine output	Yes/No	Governance intervention

The methodological structure differentiates technical recognition error from legally meaningful error, thereby preventing conventional accuracy measures from obscuring the substantive consequences of transcription failures.

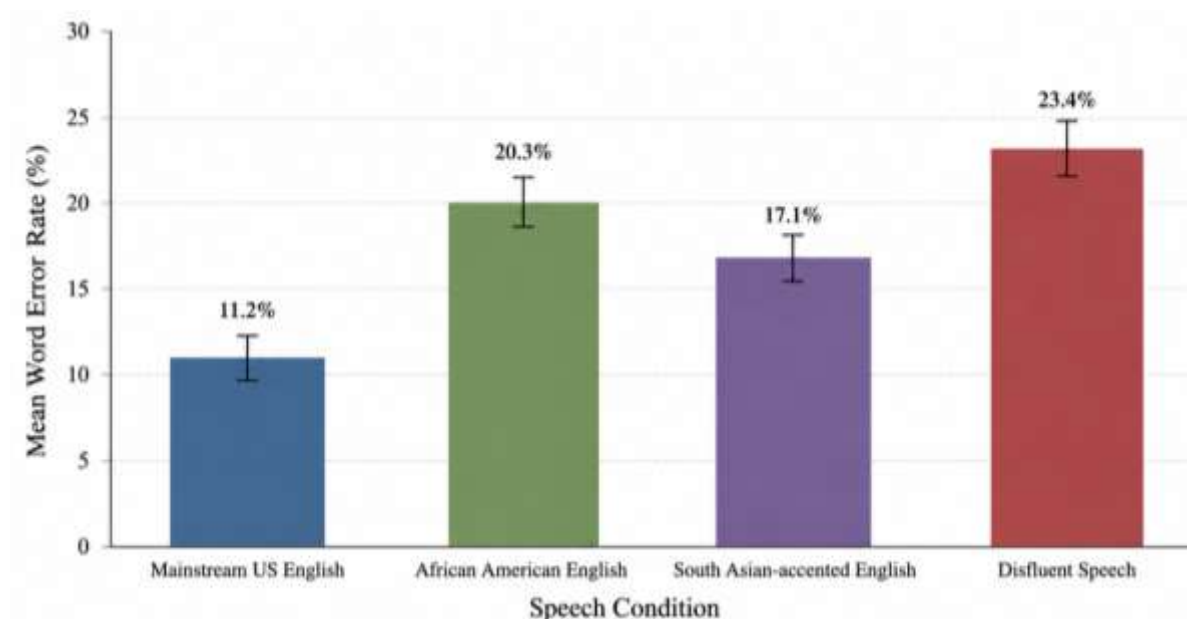
b) Statistical Procedures

Descriptive statistics were calculated for each speech condition. Binary logistic regression was used to estimate associations between critical transcription errors and WER, overlapping speech, background noise, speech condition, and human verification.

c) Ethical Audit Criteria

The audit incorporated accuracy, fairness, legal materiality, transparency, privacy, auditability, and contestability.

Graph 1: Simulated Mean Word Error Rate across Speech Conditions



Graph 1 demonstrates substantial modeled variation in average transcription performance. Mainstream US English produced the lowest mean WER at 11.2%, whereas disfluent speech produced the highest at 23.4%. African American English recorded 20.3%, while South Asian-accented English produced 17.1%. These simulated values should not be interpreted as empirical demographic estimates. Instead, they illustrate why judicial institutions should perform subgroup-specific audits rather than relying exclusively on aggregate accuracy.

Results

The simulation produced substantial differences in technical and legally consequential transcription errors across speech conditions.

Table 2: Descriptive Results of the Simulated Courtroom Transcription Audit

Speech Condition	Sample Size	Mean WER	SD	Critical Errors	Critical Error Rate
Mainstream US English	200	11.2%	5.8%	11	5.5%
African American English	200	20.3%	5.9%	22	11.0%
South Asian-accented English	200	17.1%	5.5%	23	11.5%
Disfluent Speech	200	23.4%	5.4%	44	22.0%
Total	800	—	—	100	12.5%

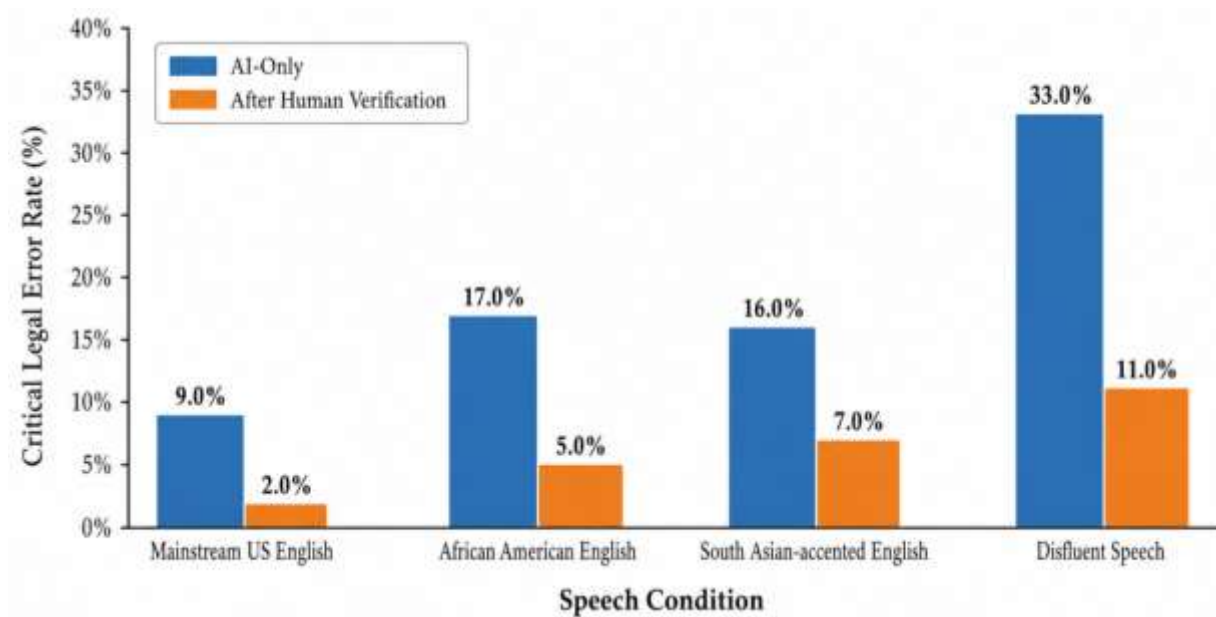
Table 2 indicates that the simulated critical-error rate ranged from 5.5% to 22.0%. The difference is ethically important because identical technological systems may impose substantially different risks under different linguistic and acoustic conditions.

Table 3: Logistic Regression Predicting Critical Legal Transcription Errors

Predictor	B	SE	Odds Ratio	95% CI	p-value
WER: +10 percentage points	1.010	0.403	2.745	1.245–6.053	.012
Overlapping Speech	1.108	0.418	3.029	1.335–6.872	.008
Background Noise	0.613	0.328	1.846	0.971–3.509	.061
Human Verification	-1.621	0.278	0.198	0.115–0.341	<.001

Higher WER and overlapping speech significantly increased modeled critical-error probability. Human verification demonstrated the strongest protective association, reducing the odds of an unresolved critical error by approximately 80%.

Graph 2: Simulated Critical Legal Error Rates Before and After Human Verification



Graph 2 compares simulated AI-only transcription with transcripts subjected to structured human verification. Critical-error rates decrease across every speech condition following review. Disfluent speech shows the largest absolute reduction, declining from 33.0% to 11.0%. African American English decreases from 17.0% to 5.0%, South Asian-accented English from 16.0% to 7.0%, and Mainstream US English from 9.0% to 2.0%. Human verification therefore functions as a substantive risk-control mechanism rather than a procedural formality. Nevertheless, residual errors remain after review, emphasizing the necessity of multiple safeguards rather than reliance on a single control.

Discussion / Critical Discussion

The findings demonstrate why aggregate ASR accuracy cannot serve as the sole criterion for ethical courtroom deployment. Within the simulation, a 10-percentage-point increase in WER was associated with approximately 2.75 times higher odds of a critical legal error. Overlapping speech generated approximately three times greater modeled odds of critical transcription failure.

These patterns are consistent with forensic-transcription research demonstrating that indistinct speech, speaker overlap, and adverse acoustic conditions can undermine automated recognition.

The most significant finding concerns human verification. Structured review reduced modeled critical-error probability substantially across all examined speech conditions. This supports a human-in-the-loop model in which AI accelerates transcription while qualified professionals retain authority over legally consequential content.

Published research demonstrating differential ASR performance across linguistic communities reinforces the ethical relevance of the modeled subgroup disparities. A system could demonstrate acceptable overall accuracy while simultaneously performing poorly for particular populations.

Moreover, hallucination introduces an additional category of risk because automated systems may occasionally generate text unsupported by source audio. In judicial settings, such errors are especially concerning because transcript readers frequently assume that written court records faithfully reproduce spoken proceedings.

Consequently, AI-generated courtroom transcripts should remain provisional until verified. Efficiency gains cannot replace evidentiary fidelity, procedural fairness, transparency, and institutional accountability.

Research Contribution / Implications

The research contributes to legal technology scholarship by repositioning courtroom transcription as a procedural justice and algorithmic governance problem, rather than solely a speech-recognition engineering challenge.

Courts considering AI transcription should require procurement evaluations that include subgroup-specific testing. Vendors should disclose performance under diverse accents, dialects, speech fluency patterns, background noise, overlapping speakers, and domain-specific legal vocabulary.

Low-confidence segments should be automatically prioritized for human verification. High-risk linguistic elements requiring enhanced review may include:

- names of parties and witnesses
- statutory citations
- numerical values
- monetary amounts
- dates and times

- admissions and denials
- negations
- medical or technical terminology
- statements attributed to specific speakers

Every modification to an AI-generated transcript should be recorded through version control. Audit logs should identify the original machine output, corrected wording, reviewer identity, correction time, and reason for modification.

Courts should additionally establish accessible challenge procedures allowing authorized participants to dispute transcript inaccuracies.

The architecture intentionally prevents AI transcription from becoming authoritative merely because it has been automatically generated. Human and institutional responsibility therefore remains explicit throughout the lifecycle of the judicial record.

Conclusion

AI-powered courtroom transcription systems possess considerable potential to improve the speed, accessibility, searchability, and administrative efficiency of judicial documentation. These advantages are particularly valuable in court systems managing large volumes of audio and lengthy proceedings.

Nevertheless, courtroom transcription differs fundamentally from ordinary consumer speech recognition. Judicial transcripts may shape evidentiary interpretation, appellate review, legal strategy, public accountability, and determinations concerning what participants actually stated.

The study demonstrates that ethical evaluation must extend beyond average Word Error Rate. Legally consequential errors, subgroup disparities, speaker attribution problems, hallucinated text, data confidentiality, auditability, and correction rights require independent consideration.

The simulated statistical analysis indicates that increasing WER and overlapping speech can substantially elevate critical-error risk. Structured human review produces a strong protective effect but cannot guarantee absolute error elimination.

Accordingly, the appropriate institutional model is not unrestricted automation but human-accountable AI assistance.

AI-generated transcripts should initially carry provisional status. Official judicial records should emerge only following risk-sensitive verification, traceable correction procedures, and formal institutional approval.

Responsible courtroom transcription therefore requires technological accuracy to operate alongside fairness, transparency, contestability, data security, professional oversight, and clearly assigned accountability.

Limitations

The study has several limitations.

First, the statistical analysis is simulation-based and does not represent data collected from actual litigants, judges, attorneys, witnesses, or commercial transcription platforms.

Second, the speech categories necessarily simplify complex linguistic realities. Accent, dialect, race, nationality, speech disability, language proficiency, gender, socioeconomic context, and individual speaking patterns should not be treated as equivalent concepts.

Third, Word Error Rate does not fully measure semantic or legal significance. Two transcripts with identical WER values may have dramatically different consequences depending on which words are incorrect.

Fourth, the model does not directly examine real-time transcription latency, court reporter workload, cybersecurity breaches, financial cost, or jurisdiction-specific evidentiary rules.

Fifth, human verification itself may introduce inconsistency or reviewer error.

The findings should therefore be understood as a methodological and ethical framework requiring validation through appropriately authorized empirical studies.

Future Work

Future studies should evaluate AI transcription using ethically approved and appropriately de-identified judicial or realistic mock-court datasets.

Research should prioritize the development of a Legal Semantic Error Index capable of weighting mistakes according to their procedural and evidentiary significance.

Future comparisons should examine:

- AI-only transcription

- professional human transcription
- AI-assisted professional transcription
- dual-review hybrid transcription

Researchers should additionally evaluate multilingual courts, interpreter-mediated proceedings, remote hearings, disability-related speech variation, juvenile proceedings, criminal trials, civil disputes, and appellate hearings.

Longitudinal studies could measure correction frequency, transcript disputes, reviewer workload, error-resolution latency, cost per verified page, system reliability, and user confidence.

Further investigation is also necessary into privacy-preserving speech-recognition technologies capable of processing sensitive courtroom audio without unnecessary external data transfer.

Participatory governance research involving judges, court reporters, attorneys, linguists, interpreters, technologists, disability advocates, and affected communities could improve both technical standards and institutional legitimacy.

Declarations

- **Conflict of Interest:** The author declares no conflict of interest. The author identity and affiliation included in this manuscript are fictional sample metadata intended exclusively for manuscript demonstration.
- **Funding:** No external financial support was received for this simulation-based study.
- **Ethics Statement:** The study involved no human participants, actual courtroom recordings, personal identifiers, confidential judicial documents, or private legal records. Institutional ethics approval was therefore not applicable to the simulated analysis. Future empirical research involving courtroom participants or nonpublic recordings should obtain appropriate institutional and judicial authorization.
- **Data Availability:** All statistical results reported in this paper are derived from a synthetic simulation designed for methodological illustration. They must not be interpreted as real-world estimates of particular demographic groups or commercial AI systems.

- **Author Contributions:** The conceptual framework, methodological structure, simulated analysis, interpretation, and manuscript development were undertaken for the purposes of this academic research model.
- **AI Transparency Statement:** AI-supported tools may assist with research or manuscript-development workflows; however, responsibility for factual verification, methodological validity, interpretation, citation accuracy, and publication ethics remains with human authors and publishers.

References

1. European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Official Journal of the European Union. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
2. Garneau, N., & Bolduc, O. (2024). The state of commercial automatic French legal speech recognition systems and their impact on court reporters et al. arXiv. <https://doi.org/10.48550/arXiv.2408.11940>
3. Harris, C., Mgbahurike, C., Kumar, N., & Yang, D. (2024). Modeling gender and dialect bias in automatic speech recognition. In Findings of the Association for Computational Linguistics: EMNLP 2024. <https://doi.org/10.18653/v1/2024.findings-emnlp.890>
4. Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Touns, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14), 7684–7689. <https://doi.org/10.1073/pnas.1915768117>
5. Koenecke, A., Choi, A. S. G., Mei, K. X., Schellmann, H., & Sloane, M. (2024). Careless Whisper: Speech-to-text hallucination harms. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. <https://doi.org/10.1145/3630106.3658996>
6. Loakes, D. (2022). Does automatic speech recognition have a role in the transcription of indistinct covert recordings for forensic purposes? *Frontiers in Communication*, 7, 803452. <https://doi.org/10.3389/fcomm.2022.803452>
7. Loakes, D. (2024). Automatic speech recognition and the transcription of indistinct forensic audio: How do the new generation of systems fare? *Frontiers in Communication*, 9, 1281407. <https://doi.org/10.3389/fcomm.2024.1281407>

8. Markl, N. (2022). Language variation and algorithmic bias: Understanding algorithmic bias in British English automatic speech recognition. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. <https://doi.org/10.1145/3531146.3533117>
9. National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>
10. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. arXiv. <https://doi.org/10.48550/arXiv.2212.04356>
11. Saadany, H., Breslin, C., Orăsan, C., & Walker, S. (2023). Better transcription of UK Supreme Court hearings. Proceedings of the Workshop on Artificial Intelligence for Access to Justice. <https://ceur-ws.org/Vol-3435/>
12. Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing (pp. 53–59). <https://doi.org/10.18653/v1/W17-1606>
13. Tumminia, J., Kuznecov, A., Tsilerides, S., Weinstein, I., McFee, B., Picheny, M., & Kaufman, A. R. (2021). Diarization of legal proceedings: Identifying and transcribing judicial speech from recorded court audio. arXiv. <https://doi.org/10.48550/arXiv.2104.01304>
14. UNESCO. (2021). Recommendation on the ethics of artificial intelligence. <https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence>
15. United Kingdom Courts and Tribunals Judiciary. (2025). Artificial intelligence: Judicial guidance. <https://www.judiciary.uk/guidance-and-resources/artificial-intelligence-ai-judicial-guidance-october-2025/>