Appendix E: Ethical Considerations and Racial Justice Act Context

CourtShadow examines institutional linguistic environments using statistical modeling. Due to the sensitive social and legal dynamics that often manifest in courtroom language, responsible interpretation is essential. This appendix discusses ethical considerations, potential risks, and the legal context provided by Racial Justice Acts (RJAs). It also clarifies the boundaries of what CourtShadow does and does not claim.

E.1 Institutional Scope and Non-Individual Interpretation

CourtShadow models *case-level environments*. It does not attempt to quantify the intentions or behavior of individual judges, prosecutors, defense attorneys, or witnesses. This distinction is in alignment with longstanding findings in sociolinguistics. For instance, institutional discourse patterns are thought to emerge from structural factors as opposed to the traits of specific participants (Conley & O'Barr, 1990; Gibbons, 2003).

CourtShadow therefore makes no claims that:

- Any individual is biased
- Any specific utterance is discriminatory
- Linguistic differences reflect intent or motive

Rather, the model detects *aggregate statistical patterns* in segment structure, framing, pronoun usage, and topic emphasis across transcripts.

E.2 Racial Justice Acts and Systemic Evidence

Recent Racial Justice Acts in California and North Carolina explicitly permit defendants to introduce statistical evidence of systemic discrimination. These may include disparities in policing, charging, jury selection, and sentencing (California Racial Justice Act, 2020; North Carolina Racial Justice Act, 2009/2020). Although RJAs do not explicitly address linguistic analysis, they acknowledge the value of *system-level patterns* in conveying deeper racial disparities.

CourtShadow is conceptually aligned with this systemic orientation:

- It focuses on *institutional discourse* instead of personal bias.
- It aggregates predictions across entire transcripts, not isolated statements.
- It provides transparent, interpretable feature contributions rather than black-box predictions.

However, CourtShadow is not designed for litigation or evidentiary use. It instead functions as exploratory analysis into the ways in which linguistic environments may correlate with group-coded categories at the corpus level.

E.3 Limitations of Observational Linguistic Data

Following best practices in computational social science and sociolegal research, several limitations constrain interpretation (Maynard, 1984; Johnson, 2020):

- **Non-causal Nature**: Logistic regression quantifies correlations in language use; it does not identify causal mechanisms behind linguistic differences.
- **Transcript Completeness**: Many transcripts omit opening statements, sidebar discussions, or pretrial hearings. These components can shift linguistic distributions.
- **Case-type Heterogeneity**: Although topic indicators partially address differences, variance in charges, witnesses, and trial phases remains a confounder.
- **Role Distributions**: Judges, attorneys, and witnesses contribute differently to the linguistic environment; unequal role representation may influence predictions.
- **Archival Bias**: Transcripts available to the public may differ systematically from those not released or pay-walled.

These limitations motivate CourtShadow's emphasis on transparency and interpretability as opposed to solely relying on predictive performance.

E.4 Avoiding Overinterpretation

Logistic regression decomposes contributions linearly, meaning that it tends to overemphasize the importance of specific words or features. CourtShadow is designed to counteract this tendency:

- Feature-family decomposition highlights *patterns*, not keywords
- Case-level averaging reduces the influence of outlier segments
- Calibration checks ensure that probabilities reflect appropriate uncertainty

CourtShadow should therefore be treated as an analytic lens—not as diagnostic evidence about any individual, case, or legal outcome.

E.5 Responsible Use Principles

The following principles guide appropriate interpretation and deployment:

1. **Transparency**: All features are deliberately interpretable, and all modeling steps are documented.
2. **Contextualization**: Results should be interpreted alongside qualitative understanding of courtroom dynamics (Conley & O'Barr, 1990; Gibbons, 2003).
3. **Caution**: Model outputs should not be used for any legal decision-making, classification of individuals, or normative judgments.
4. **Boundary awareness**: CourtShadow analyzes *linguistic environments*. Racial identity, criminal behavior, or judicial fairness are not analyzed directly.
5. **Ethical alignment**: The project mirrors RJA frameworks by focusing on systemic patterns as opposed to than individual blame.

E.6 Summary

CourtShadow's purpose is to surface linguistic patterns in courtroom discourse via a transparent statistical model. It contributes to research on institutional communication, sociolinguistic inequality, and the emerging legal interest in systemic evidence. Importantly, it CourtShadow makes no normative or causal claims. When studying language in legal settings, there are numerous ethical constraints that arise. All results from this tool must be contextualized within such constraints and interpreted with caution.