Supplementary Materials for

Ethnic Bias in Judicial Decision-making: Evidence from Criminal Appeals in Kenya

Donghyun Danny Choi, J. Andrew Harris, Fiona Shen-Bayh

Supplementary Appendix

A Data

We assembled data on appeals from the Kenya Law Cases Database, an online repository of court rulings maintained by the National Council for Law Reporting (Kenya Law).[33] The full texts of court rulings are available in XML format for individual download. Using a web driver, we crawled and downloaded the XML files of 9,545 criminal appeals rulings issued by the High Court between January 1, 2003 and December 31, 2017. We then used regular expression text extractions to identify critical case features.

For each criminal appeal, we used XML header tags to extract the date of ruling, case number, and county court where the case was heard. We also identified the names of the appellant and respondent from the case citation where the appellant is always listed before the respondent, e.g. Mithungi [appellant] v. Republic [respondent]. To extract the names of individual judges, we utilized the fact that judge names are typically located near the end of a ruling, often following the word "Judge." Furthermore, when multiple judges vote together in a case, their names are located next to one another in the text. We leveraged these spatial patterns to extract 133 unique judges.

We then used the names of appellants to estimate their ethnic identity. Given the relatively limited number of judges in the data, a member of the Kenyan legal community coded ambiguous judges ethnicities by canvassing their professional networks to learn about the judges' ethnicity. In contrast, names for the appellants and respondents were too numerous to code by hand. To solve this problem, we build upon data and methods in Harris (2015) to estimate the ethnicity of participants in legal proceedings. Our approach leverages information from Kenya's voter register, which identifies voter names from ethnically homogeneous areas. We use this data to create a dictionary-based ethnicity classifier that estimates the probability of ethnicity for a given name. Then, we use these probabilities to link each person's name to an ethnic group.[34]

Extracting the final outcome of each appeal presented complications. The language of appeal decisions does not always follow a consistent pattern. While some phrases remain relatively constant (e.g. "This court hereby finds..."), patterns of judicial speech sometimes vary considerably by judge and year. Furthermore, some texts feature summaries of a ruling made in a previous case, using language that might be captured by our regular expressions extractor and thus incorrectly classified as a final appeal outcome. To address these concerns, we compiled two mutually exclusive lists of expressions designed to capture appeals that were either allowed or denied.[35] We also utilized the fact

[33] Originally established by The National Council for Law Reporting Act (1994), Kenya Law Cases is the most comprehensive legal database in Kenya. It contains the full text decisions of civil and criminal cases delivered by magistrate courts, High Courts, the Appeals Court, the Supreme Court, and other special tribunals.

[34] For simplicity, we include 12 (of ~ 42) ethnic groups in Kenya: Kalenjin, Kamba, Kikuyu, Kisii, Luhya, Luo, Masai, Meru, Mijikenda, Pokot, Somali, and Turkana. We make this simplification for two reasons. These 12 groups make up over 90% of the population. And for many of the smaller ethnic groups, a lack of group-specific naming conventions or simply a small numbers of individuals means that they are often indistinguishable from other nearby groups, or intermixed in a way that makes a name-based approach to identity futile.

[35] Successful appeals tend to feature words such as "allowed", "succeeds", "finds merit", and "conviction overturned", while rejected appeals tend to feature phrases such as "is denied", "is dismissed", "is rejected", "finds no merit", and "conviction upheld". These phrases represent a small subset of the full list used (see appendix).

that previous rulings tended to be summarized in the opening paragraph of each decision, whereas final outcomes tended to appear in the closing paragraph. By restricting our text extractor to the last few paragraphs of each decision, we thus limited misclassification of previous rulings as final outcomes.[36] Taking these steps enabled us to code the outcomes for approximately 80% the appeals in our sample.[37]

Finally, we used regular expressions to extract details related to whether the defendant was convicted of crimes relating to persons, property, health, or morality,[38] as well as whether they were sentenced to death, imprisonment, or corporal punishment. Details about the original case were often featured in the opening lines of the decision, as described above. We converted these crime and sentencing features into fixed effects in the analyses that follow.

B Balance Tests

Table B1: Balance Tests

Term	Estimate	Std.Error	T Statistic	P-Value
Case Type				
Murder	-0.025	0.015	-1.739	0.083
Manslaughter	0.046	0.020	2.233	0.026
Violence	0.007	0.007	0.893	0.373
Vehicle	0.020	0.013	1.480	0.140
Arson	-0.004	0.011	-0.367	0.714
Drug	-0.002	0.015	-0.151	0.880
Theft	-0.017	0.008	-2.110	0.036
Public order	0.033	0.031	1.073	0.284
Prior Sentence				
Death	-0.027	0.018	-1.451	0.148
Prison	0.002	0.009	0.251	0.802
Stroke	0.006	0.026	0.215	0.830

[36] There was an important trade-off between the size of the text window and classification accuracy. Shrinking the text window reduced the number of false classifications, but at the risk of truncating relevant information about case outcomes and producing null results. Expanding the text window raised the number of positive classifications, but at the risk of capturing information about a previous case rather than a final outcome. To account for this trade-off, the window size was manually adjusted for each year in the sample (2003-2017) in order to minimize the number of null results and false positives.

[37] The remaining 20% either had outcomes that were contained earlier in the text (which was removed by our length restrictions) or used idiosyncratic or misspelled language that were not captured by regular expressions.

[38] These classifications are based on the Kenyan Penal Code.

C Dictionaries

A key consideration of any dictionary method is which words best represent our sentiment of interest. Yet, word selection is often an arbitrary process and many studies do not offer guidance on optimal selection criteria (Grimmer and Stewart, 2013). Recent advances in natural language processing reveal how word embeddings can be used to populate dictionaries with minimal supervision.[39] The intuition of this approach is that every word in a corpus can be represented as a vector, the mapping of which can be interpreted as a spatial representation of the sentiment of that word; words that are closer together in the same vector space can thus be thought of as syntactically similar (Pennington, Socher and Manning, 2014). Furthermore, by vectorizing the vocabulary, the semantic similarity between any two words can be quantified using vector operations, specifically the cosine of the angle between two word vectors. Leveraging these properties, we build a sentiment dictionary with a small set of seed words and using cosine similarity scores to populate each dictionary with a list of most-similar terms.

Given the trade-offs of different dictionary approaches,[40] we evaluate sentiment using two sets of dictionaries: one derived from the corpus itself and another derived from an off-the-shelf vocabulary.[41] The findings from the corpus-derived dictionaries our featured in our main results section; findings from the off-the-shelf dictionaries, which are consistent with our main results, are saved for the Appendix (see Robustness: Text Analysis).

Our main text analysis uses a word embeddings model to build a minimally supervised dictionary for each sentiment of interest. For sentiment category s, we randomly sampled three seed words from our lists of trust- and disgust-related terms. We then used these seed words to retrieve the 2000 most-similar words from the corpus, where similarity was calculated using the cosine of the angle between two word vectors. Our word vectors were derived using the Global Word Vectors (GloVe) model. Because word embeddings require a vast amount of training data in order to produce stable vector representations (Antoniak and Mimno, 2018), we used the Common Crawl GLoVe model that was trained using a 1.9 million word vocabulary. Rodriguez and Spirling (N.d.) further find that pretrained word vectors perform well against both locally trained vectors and human coders.

Using the embeddings procedure described above, we derived sentiment dictionaries as shown in Table C1 where each column shows the most-similar terms derived from a set of sentiment seed words. The top three rows are the three randomly selected seed words from each category; the bottom rows show the top ten most similar words, where similarity is calculated using the cosine similarity scores derived from the GloVe model. We show the embeddings-derived dictionaries for the off-the-shelf seed terms below under Robustness: Text Analysis.

[39] Our approach builds on the work of Rice and Zorn (2019), who use such techniques to develop polarity scores measuring the proportion difference in opposing sentiments (e.g., positive versus negative).

[40] Corpus-specific dictionaries are better attuned to how keywords are used in context, but can be more difficult to validate (Loughran and McDonald, 2011); off-the-shelf dictionaries are more comprehensive, but more general vocabularies can be misapplied to niche corpora (Rice and Zorn, 2019; Grimmer and Stewart, 2013).

[41] The NRC Emotion Lexicon. See https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.

Table C1: Seeds and Retrieved Words for Sentiment Categories: Corpus-Derived Words

	Trust	Disgust	Positive	Negative
Seed Words	trustworthiness	cruel	positive	negative
	honest	heinous	good	bad
	confident	immoral	correct	incorrect
Retrieved (Top 10)	truthful	brutal	better	wrong
	confidence	inhuman	should	worse
	respectful	shameful	possible	unfortunately
	legitimate	evil	kind	mistaken
	genuinely	stupid	right	problem
	caring	vicious	proper	poor
	reliable	ridiculous	fine	lack
	assured	absurd	truth	worst
	competent	dishonest	consistently	misleading
	integrity	irresponsible	correctly	avoid

D Robustness: Judgment for Defendant

In this appendix, we consider two extensions of our analyses in order to assess the robustness of our results. First, we use the ethnic profiles of judges at court-stations in given years to estimate the probability of treatment – being matched to a coethnic judge – for each appellant. Second, we consider the robustness of our results taking into account estimation uncertainty related to the appellant ethnicity categorization.

D.1 Inverse Probability Weighting

In this appendix subsection, we replicate the main results using inverse probability weighting (IPW), following the discussion in Hernan and Robins (2020, p. 23). The logic behind IPW lies in adjusting our sample to account for varying probabilities of treatment which might bias our estimates of treatment, effectively reweighting the sample. For a given probability of treatment p_i for unit i, the inverse probability of treatment weight for a treated unit is $\frac{1}{p_i}$ and for a control unit is $\frac{1}{1-p_i}$.

In many applications, these probabilities are estimated with covariates using standard propensity score techniques. In our case, we can approximate the *actual* probability of treatment for each appellant by calculating the distribution of judge ethnicities at the court station where the appeal was filed. This is important because, conditional on the appellant, the distribution of ethnicities of available judges defines the probability of treatment for that appellant at a given station. For instance, a Kikuyu appellant faced with a pool of judges – a Kikuyu, two Luos, and a Mijikenda – has a probability of treatment equal to 0.25.

Inverse probability weighting has an appealing property, in that it allows us to identify and drop observations that violate the positivity assumption (e.g., observations that have probability of treatment equal to zero or one.) Such observations are not properly randomized, since they will always (or never) receive treatment. In our application, this is particularly pertinent, since the ethnic geography of Kenya means that certain court stations will likely hear cases from appellants from the locally predominant ethnic group. If the judges hearing cases all come from that same locally predominant ethnic group, then many appellants will face a probability of treatment equal to one. The cost is sample size: units that never or always receive treatment are dropped from the analysis.

Given that we have no information about when an appeal was filed and assigned to a judge, we cannot know for certain the *precise* distribution of judge ethnicities at a given court station. We start by approximating the probability of treatment for each appellant by calculating the distribution of judge ethnicities at each court station in the year in which the appellant's judgment is delivered. Given case backlogs, the length of time it takes to obtain a judgment, and rotation of judges in and out of a court station over time, it is likely that the actual distribution of judge ethnicities when the case was filed/assigned would be different from that distribution when the judgment was delivered. Thus, in addition to calculating this distribution using the set of judges at station j in year t, we also calculate that distribution using judges from years $t-1$ and t (table xx), $t-2$ and t (table xx), $t-3$ and t (table xx), and $t-4$ and t (table xx) below. In short, we examine how the results change across a range of plausible IPW's derived from reasonable approximations of appellants' individual probabilities of treatment. We find positive point estimates that largely comport with the main results in the text. Moreover, as the timespan of the judge-ethnicity distribution increases, we discard fewer units due to the positivity issue discussed above, and statistical significance approaches that of the primary results presented in main text.

D.2 Replication of Table 1 with IPW

Table D1: Coethnic Bias in Criminal Appeal Decisions – 1-year IPW

	Outcome: Judgement for the Defendant					
	(1)	(2)	(3)	(4)	(5)	(6)
Coethnic Match	0.046***	0.048***	0.040*	0.041*	0.037	0.030
	(0.018)	(0.018)	(0.022)	(0.021)	(0.028)	(0.029)
Courthouse-Year FE	No	No	Yes	Yes	Yes	Yes
Individual Judge FE	No	No	No	No	Yes	Yes
Case-specific Controls	No	Yes	No	Yes	No	Yes
Observations	3,008	3,008	3,008	3,008	3,008	3,008
R^2	0.002	0.015	0.108	0.119	0.153	0.161

Note: *p<0.1; **p<0.05; ***p<0.01. Coefficients estimated using OLS with one-year inverse probability weights. "Coethnic Match" is a binary variable equal to one if the judge and appellant share the same ethnic group, zero otherwise.

Table D2: Coethnic Bias in Criminal Appeal Decisions – 2-year IPW

	Outcome: Judgement for the Defendant					
	(1)	(2)	(3)	(4)	(5)	(6)
Coethnic Match	0.046***	0.045***	0.043*	0.044**	0.036	0.031
	(0.016)	(0.016)	(0.022)	(0.021)	(0.026)	(0.026)
Courthouse-Year FE	No	No	Yes	Yes	Yes	Yes
Individual Judge FE	No	No	No	No	Yes	Yes
Case-specific Controls	No	Yes	No	Yes	No	Yes
Observations	3,863	3,863	3,863	3,863	3,863	3,863
R^2	0.002	0.014	0.104	0.113	0.143	0.150

Note: *p<0.1; **p<0.05; ***p<0.01. Coefficients estimated using OLS with two-year inverse probability weights. "Coethnic Match" is a binary variable equal to one if the judge and appellant share the same ethnic group, zero otherwise.

Table D3: Coethnic Bias in Criminal Appeal Decisions – 3-year IPW

	Outcome: Judgement for the Defendant					
	(1)	(2)	(3)	(4)	(5)	(6)
Coethnic Match	0.052***	0.049***	0.058***	0.059***	0.054**	0.048*
	(0.015)	(0.015)	(0.022)	(0.021)	(0.027)	(0.026)
Courthouse-Year FE	No	No	Yes	Yes	Yes	Yes
Individual Judge FE	No	No	No	No	Yes	Yes
Case-specific Controls	No	Yes	No	Yes	No	Yes
Observations	4,436	4,436	4,436	4,436	4,436	4,436
R^2	0.003	0.015	0.105	0.114	0.141	0.149

Note: *p<0.1; **p<0.05; ***p<0.01. Coefficients estimated using OLS with three-year inverse probability weights. "Coethnic Match" is a binary variable equal to one if the judge and appellant share the same ethnic group, zero otherwise.

Table D4: Coethnic Bias in Criminal Appeal Decisions – 4-year IPW

	Outcome: Judgement for the Defendant					
	(1)	(2)	(3)	(4)	(5)	(6)
Coethnic Match	0.052***	0.048***	0.057**	0.057**	0.052**	0.047*
	(0.014)	(0.014)	(0.023)	(0.022)	(0.025)	(0.025)
Courthouse-Year FE	No	No	Yes	Yes	Yes	Yes
Individual Judge FE	No	No	No	No	Yes	Yes
Case-specific Controls	No	Yes	No	Yes	No	Yes
Observations	4,734	4,734	4,734	4,734	4,734	4,734
R^2	0.003	0.016	0.105	0.115	0.141	0.150

Note: *p<0.1; **p<0.05; ***p<0.01. Coefficients estimated using OLS with four-year inverse probability weights. "Coethnic Match" is a binary variable equal to one if the judge and appellant share the same ethnic group, zero otherwise.

Table D5: Coethnic Bias in Criminal Appeal Decisions – 5-year IPW

	Outcome: Judgement for the Defendant					
	(1)	(2)	(3)	(4)	(5)	(6)
Coethnic Match	0.048***	0.043***	0.057**	0.056**	0.050**	0.046*
	(0.014)	(0.014)	(0.023)	(0.023)	(0.024)	(0.024)
Courthouse-Year FE	No	No	Yes	Yes	Yes	Yes
Individual Judge FE	No	No	No	No	Yes	Yes
Case-specific Controls	No	Yes	No	Yes	No	Yes
Observations	4,960	4,960	4,960	4,960	4,960	4,960
R^2	0.002	0.016	0.104	0.115	0.140	0.150

Note: *p<0.1; **p<0.05; ***p<0.01. Coefficients estimated using OLS with five-year inverse probability weights. "Coethnic Match" is a binary variable equal to one if the judge and appellant share the same ethnic group, zero otherwise.

D.3 Replication of Table 2 with IPW

Table D6: Effect of Coethnic Match between Appellant and Judge, by Judge Ethnicity and One-year IPW.

	Dependent variable:						
	Outcome: Judgement for the Defendant						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Coethnic Match	0.047	−0.013	0.001	0.106	−0.055	0.028	0.065
	(0.051)	(0.045)	(0.040)	(0.081)	(0.128)	(0.156)	(0.245)
Sample	Kikuyu	Kalenjin	Luhya	Luo	Kamba	Kisii	Other
Courthouse-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual Judge FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Case-specific Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	691	335	984	406	190	131	271
R^2	0.209	0.263	0.208	0.175	0.196	0.409	0.264

Note: *p<0.1; **p<0.05; ***p<0.01. Coefficients estimated using OLS with one-year inverse probability weights. "Coethnic Match" is a binary variable equal to one if the judge and appellant share the same ethnic group, zero otherwise.

Table D7: Effect of Coethnic Match between Appellant and Judge, by Judge Ethnicity with Two-year IPW.

	Dependent variable:						
	Outcome: Judgement for the Defendant						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Coethnic Match	0.056	0.023	−0.014	0.072	−0.020	0.001	−0.022
	(0.038)	(0.049)	(0.044)	(0.064)	(0.077)	(0.145)	(0.263)
Sample	Kikuyu	Kalenjin	Luhya	Luo	Kamba	Kisii	Other
Courthouse-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual Judge FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Case-specific Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	822	420	1,314	566	250	151	340
R^2	0.207	0.282	0.177	0.147	0.166	0.397	0.259

Note: *p<0.1; **p<0.05; ***p<0.01. Coefficients estimated using OLS with two-year inverse probability weights. "Coethnic Match" is a binary variable equal to one if the judge and appellant share the same ethnic group, zero otherwise.

Table D8: Effect of Coethnic Match between Appellant and Judge, by Judge Ethnicity with Three-Year IPW.

	Dependent variable:						
	Outcome: Judgement for the Defendant						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Coethnic Match	0.095**	0.063	0.022	0.064	−0.040	−0.006	0.030
	(0.042)	(0.037)	(0.044)	(0.067)	(0.070)	(0.155)	(0.233)
Sample	Kikuyu	Kalenjin	Luhya	Luo	Kamba	Kisii	Other
Courthouse-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual Judge FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Case-specific Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	935	504	1,511	653	291	176	366
R^2	0.210	0.277	0.169	0.143	0.148	0.396	0.277

Note: *p<0.1; **p<0.05; ***p<0.01. Coefficients estimated using OLS with three-year inverse probability weights. "Coethnic Match" is a binary variable equal to one if the judge and appellant share the same ethnic group, zero otherwise.

Table D9: Effect of Coethnic Match between Appellant and Judge, by Judge Ethnicity with Four-Year IPW.

	Dependent variable:						
	Outcome: Judgement for the Defendant						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Coethnic Match	0.107***	−0.034	0.022	0.054	−0.006	−0.002	0.032
	(0.036)	(0.042)	(0.045)	(0.066)	(0.056)	(0.140)	(0.229)
Sample	Kikuyu	Kalenjin	Luhya	Luo	Kamba	Kisii	Other
Courthouse-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual Judge FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Case-specific Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,003	588	1,552	700	308	185	398
R^2	0.221	0.255	0.168	0.131	0.148	0.405	0.285

Note: *p<0.1; **p<0.05; ***p<0.01. Coefficients estimated using OLS with four-year inverse probability weights. "Coethnic Match" is a binary variable equal to one if the judge and appellant share the same ethnic group, zero otherwise.

Table D10: Effect of Coethnic Match between Appellant and Judge, by Judge Ethnicity with Five-Year IPW.

	Dependent variable:						
	Outcome: Judgement for the Defendant						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Coethnic Match	0.093**	−0.011	0.027	0.054	−0.001	−0.040	0.050
	(0.034)	(0.030)	(0.045)	(0.063)	(0.052)	(0.146)	(0.232)
Sample	Kikuyu	Kalenjin	Luhya	Luo	Kamba	Kisii	Other
Courthouse-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual Judge FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Case-specific Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,038	645	1,568	755	332	191	431
R^2	0.220	0.242	0.169	0.143	0.145	0.407	0.294

Note: *p<0.1; **p<0.05; ***p<0.01. Coefficients estimated using OLS with five-year inverse probability weights. "Coethnic Match" is a binary variable equal to one if the judge and appellant share the same ethnic group, zero otherwise.

D.4 Dropping "Lost Interest" Cases

Tables D11 and D12 drop observations that represent administrative clearance of appeals that had not seen regular activity in several years. These 45 cases represent situations where the judge asserts that, given the lack of activity from the appellant's side, the appellant has lost interest in pursuing the appeal. When these cases are dropped, the results marginally strengthen.

Table D11: Effect of Coethnic Match, Dropping "Lost Interest" Cases

	Outcome: Judgement for the Defendant					
	(1)	(2)	(3)	(4)	(5)	(6)
Coethnic Match	0.044***	0.041***	0.043**	0.042**	0.039***	0.036**
	(0.015)	(0.015)	(0.018)	(0.017)	(0.014)	(0.014)
Courthouse-Year FE	No	No	Yes	Yes	Yes	Yes
Individual Judge FE	No	No	No	No	Yes	Yes
Case-specific Controls	No	Yes	No	Yes	No	Yes
Observations	9,500	9,500	9,500	9,500	9,500	9,500
R^2	0.001	0.008	0.078	0.084	0.104	0.108

Note: *p<0.1; **p<0.05; ***p<0.01. Coefficients estimated using OLS. "Coethnic Match" is a binary variable equal to one if the judge and appellant share the same ethnic group, zero otherwise. Excludes cases dropped due to loss of interest from appellant.

Table D12: Effect of Coethnic Match, Dropping "Lost Interest" Cases, by Judge Ethnicity

	Dependent variable:						
	Outcome: Judgement for the Defendant						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Coethnic Match	0.060**	0.025	−0.018	0.024	0.079**	0.096	0.057
	(0.023)	(0.022)	(0.029)	(0.061)	(0.029)	(0.130)	(0.094)
Sample	Kikuyu	Kalenjin	Luhya	Luo	Kamba	Kisii	Other
Courthouse-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual Judge FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Case-specific Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,192	1,042	2,916	1,216	760	531	843
R^2	0.164	0.146	0.132	0.091	0.087	0.223	0.203

Note: *p<0.1; **p<0.05; ***p<0.01. Coefficients estimated using OLS with five-year inverse probability weights. "Coethnic Match" is a binary variable equal to one if the judge and appellant share the same ethnic group, zero otherwise. Excludes cases dropped due to loss of interest from appellant.

D.5 Accounting for judges who were deemed unsuitable for higher positions in the judiciary due to corruption

Table D13: Effect of Coethnic Match with Controls for Judges Ineligible for Promotion

	Outcome: Judgement for the Defendant					
	(1)	(2)	(3)	(4)	(5)	(6)
Coethnic Match	0.042***	0.039***	0.043**	0.041**	0.036**	0.033**
	(0.015)	(0.015)	(0.018)	(0.018)	(0.014)	(0.014)
Courthouse-Year FE	No	No	Yes	Yes	Yes	Yes
Individual Judge FE	No	No	No	No	Yes	Yes
Case-specific Controls	No	Yes	No	Yes	No	Yes
Observations	9,545	9,545	9,545	9,545	9,545	9,545
R^2	0.001	0.009	0.079	0.085	0.105	0.110

Note: *p<0.1; **p<0.05; ***p<0.01. Coefficients estimated using OLS. "Coethnic Match" is a binary variable equal to one if the judge and appellant share the same ethnic group, zero otherwise. Specifications include dichotomous variables for judges declared ineligible for higher judicial posts by the Judicial Service Commission review.

Table D14: Effect of Coethnic Match Omitting Judges Ineligible for Promotion

	Outcome: Judgement for the Defendant					
	(1)	(2)	(3)	(4)	(5)	(6)
Coethnic Match	0.048***	0.044***	0.048**	0.046**	0.037**	0.034**
	(0.015)	(0.015)	(0.019)	(0.018)	(0.015)	(0.015)
Courthouse-Year FE	No	No	Yes	Yes	Yes	Yes
Individual Judge FE	No	No	No	No	Yes	Yes
Case-specific Controls	No	Yes	No	Yes	No	Yes
Observations	9,267	9,267	9,267	9,267	9,267	9,267
R^2	0.001	0.010	0.080	0.086	0.104	0.109

Note: *p<0.1; **p<0.05; ***p<0.01. Coefficients estimated using OLS. "Coethnic Match" is a binary variable equal to one if the judge and appellant share the same ethnic group, zero otherwise. Specifications exclude judges declared ineligible for higher judicial posts by the Judicial Service Commission review.

D.6 Alternative Aggregation of Ethnic Groups

Table D15: Effect of Coethnic Match between Appellant and Judge.

	Outcome: Judgement for the Defendant					
	(1)	(2)	(3)	(4)	(5)	(6)
Coethnic Match	0.041***	0.039***	0.043**	0.042**	0.037**	0.035**
	(0.015)	(0.015)	(0.018)	(0.017)	(0.014)	(0.014)
Courthouse-Year FE	No	No	Yes	Yes	Yes	Yes
Individual Judge FE	No	No	No	No	Yes	Yes
Case-specific Controls	No	Yes	No	Yes	No	Yes
Observations	9,545	9,545	9,545	9,545	9,545	9,545
R^2	0.001	0.009	0.079	0.085	0.105	0.110

Note: *p<0.1; **p<0.05; ***p<0.01. Coefficients estimated using OLS. "Coethnic Match" is a binary variable equal to one if the judge and appellant share the same ethnic group, zero otherwise. Uses more aggregated version of ethnic match variable.

Table D16: Effect of Coethnic Match between Appellant and Judge, by Judge Ethnicity.

	Dependent variable:						
	Outcome: Judgement for the Defendant						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Coethnic Match	0.057**	0.054	−0.018	0.024	0.079**	0.096	0.042
	(0.023)	(0.032)	(0.029)	(0.061)	(0.029)	(0.130)	(0.083)
Sample	Kikuyu	Kalenjin	Luhya	Luo	Kamba	Kisii	Other
Courthouse-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual Judge FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Case-specific Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,235	1,042	2,917	1,217	760	531	843
R^2	0.169	0.147	0.132	0.091	0.087	0.223	0.203

Note: *p<0.1; **p<0.05; ***p<0.01. Coefficients estimated using OLS. "Coethnic Match" is a binary variable equal to one if the judge and appellant share the same ethnic group, zero otherwise. Uses more aggregated version of ethnic match variable.

D.7 Uncertainty over appellant ethnicity

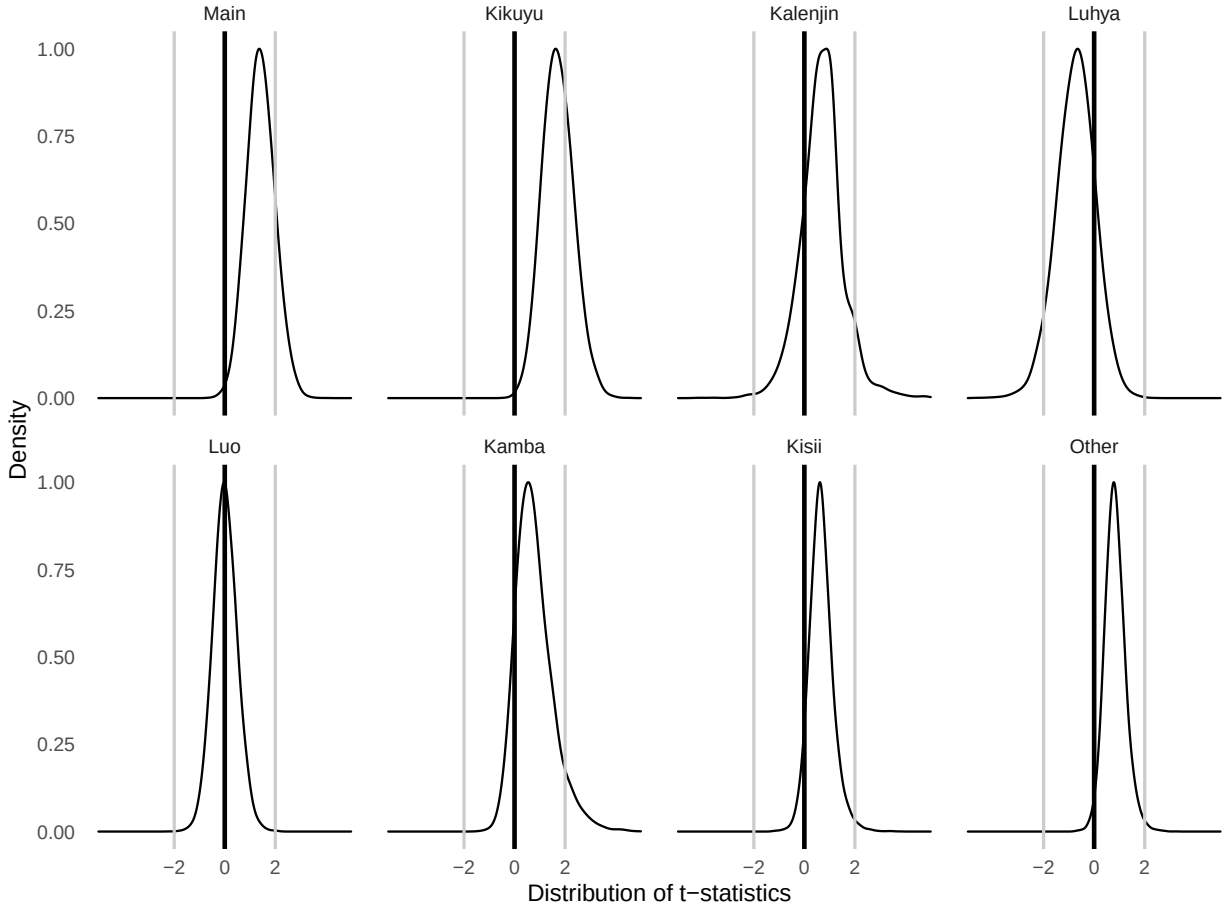
In the main text, we take the estimated highest probability ethnic group associated with the appellant, and classify that group as the appellant's group. This approach underestimates the measurement uncertainty, since there is some chance that the appellant could identify with another group. In this subsection, we simulate this uncertainty to see how it affects the primary results. Our expectation is that the results will retain the observed sign but, given the additional noise, the results will attenuate.

To simulate, we proceed as follows for 10000 iterations.

1. For appellant i, calculate the probability of membership to each ethnic group g.
2. Draw a random variable from the distribution defined by the calculated categorical probabilities. This draw represents one possible ethnic assignment for the appellant.
3. Create the ethnic match variable based on this random draw for the appellant.
4. Estimate regression coefficients based on this new ethnic match variable.
5. Store the coefficients and standard errors and repeat.

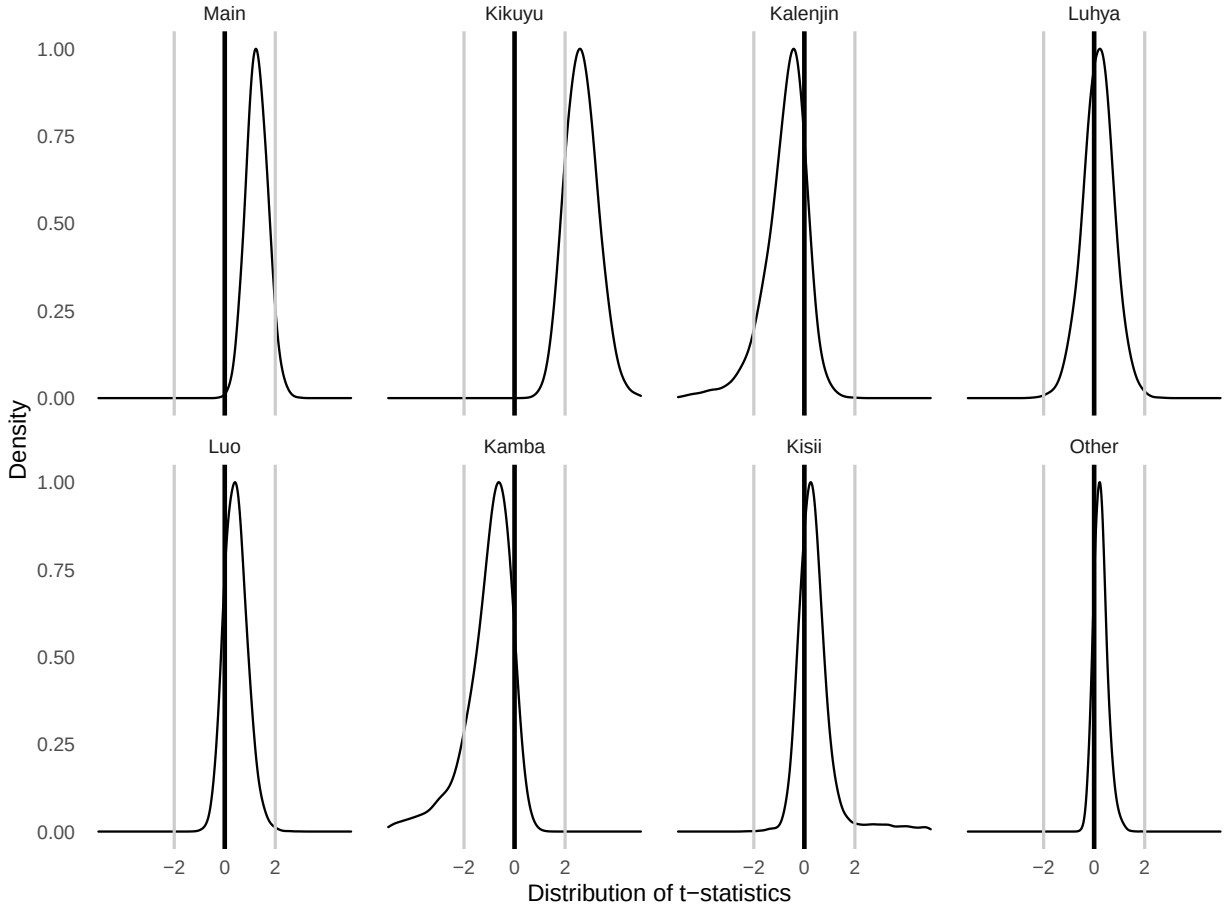
Figure D1 presents the distribution of t-statistics from these 80000 regressions (10000 x 8 table columns) for the rightmost column in table 1 (containing the most stringent fixed effect specification) and all columns of table 3 (by judge-ethnic group). The figure largely supports the main results. The modes for both the main result in table 1 and the result from table 3 on Kikuyu judges remain near the deterministic results, while the other analyses remain mostly below standard levels of statistical significance. Figure D2 replicates Figure D1 using the four-year inverse-probability weights.

Figure D1: Appeal Outcome: Robustness Check



Notes: The figure shows the distribution of t-statistics for each regression model, where the match variable incorporates uncertainty about appellant ethnicity by randomly selecting appellant ethnicity from the estimated distribution across ethnic groups.

Figure D2: Appeal Outcome: Robustness Check, Four-Year IPW.



Notes: The figure shows the distribution of t-statistics for each regression model, where the match variable incorporates uncertainty about appellant ethnicity by randomly selecting appellant ethnicity from the estimated distribution across ethnic groups.

E Robustness: Text Analysis

We validated our main text analysis two ways: first, we ran a GloVe model on our corpus to derive a corpus-specific set of word embeddings; second, we created a second set of dictionaries using the NRC Emotion Lexicon. We then evaluated our analysis using every possible combination of vectors and dictionaries – pretrained vectors, corpus-trained vectors, corpus seeds, NRC seeds. Table E1 reveals that our main results are unchanged using NRC seed words, wherein judges use approximately 7% more trustworthy terms when hearing the case of a coethnic appellant than a non-coethnic; this estimate is statistically significant at the 5% confidence level. This relationship appears even stronger when using vectors modeled directly from the corpus, as shown in Table E2 – the coethnic match variable is now approximately 12% at the 1% confidence level. Our results are consistent when we combine corpus-trained vectors with corpus-derived seed words, as shown in Table E3, where the coethnic match variable is approximately 11% and still statistically significant at the 1% confidence level. The variability of our results with the corpus-trained vectors is in line with broader research on the challenges of running such models on smaller bodies of text. As Antoniak and Mimno (2018) observe, cosine similarity scores are more volatile when trained on relatively small, niche corpora. It thus makes sense that our findings from the pretrained GloVe vectors – estimated using nearly 2 billion words – are considerably more stable.

Table E1: Coethnic Bias in Written Judgments: NRC Seeds, GloVe Vectors

	Dependent variable:			
		Sentiment		
	Trust	Disgust	Positive	Negative
	(1)	(2)	(3)	(4)
Coethnic Match	0.074**	0.022	0.082**	0.011
	(0.031)	(0.038)	(0.037)	(0.033)
Individual Judge FE	Yes	Yes	Yes	Yes
Courthouse-Year FE	Yes	Yes	Yes	Yes
Observations	9,545	9,545	9,545	9,545
R^2	0.235	0.214	0.221	0.238
Note:				*p<0.1; **p<0.05; ***p<0.01

Table E2: Coethnic Bias in Written Judgments: NRC Seeds, Corpus-Derived Vector

	Dependent variable:			
	Sentiment			
	Trust	Disgust	Positive	Negative
	(1)	(2)	(3)	(4)
Coethnic Match	0.122***	0.052*	0.022	0.022
	(0.040)	(0.027)	(0.035)	(0.036)
Individual Judge FE	Yes	Yes	Yes	Yes
Courthouse-Year FE	Yes	Yes	Yes	Yes
Observations	9,545	9,545	9,545	9,545
R^2	0.194	0.233	0.168	0.227
Note:				*p<0.1; **p<0.05; ***p<0.01

Table E3: Coethnic Bias in Written Judgments: Corpus Seeds, Corpus-Derived Vectors

	Dependent variable:			
	Sentiment			
	Trust	Disgust	Positive	Negative
	(1)	(2)	(3)	(4)
Coethnic Match	0.111***	0.045	0.027	0.024
	(0.035)	(0.036)	(0.036)	(0.029)
Individual Judge FE	Yes	Yes	Yes	Yes
Courthouse-Year FE	Yes	Yes	Yes	Yes
Observations	9,545	9,545	9,545	9,545
R^2	0.154	0.165	0.241	0.178
Note:				*p<0.1; **p<0.05; ***p<0.01

F Robustness: Validation of Word Embeddings

In this appendix, we report on a validation exercise designed to evaluate the degree to which the word embeddings we employ[42] approximate human usage of words.

Word embeddings systematically encode how words relate to one another. The degree to which word embeddings approximate human use of words is an open question within any given application. We use the Turing test described in Rodriguez and Spirling (N.d.) to investigate this question. In general, a Turing test is an "imitation game" (Turing, 1950). A human is confronted with two signals – sentences, for instance – one generated by a human and the other generated by a computer. If the human is unable to distinguish which signal was generated by which source, then the computer has passed the Turing test.

In our case, word embeddings take the role of the computer. In the first step, the word embeddings are used to retrieve ten context words most similar to various legally-relevant cue words.[43] Similarly, we task a set of mTurk workers with listing their top ten context words for the same cue words, and taking the top ten words most frequently mentioned by these workers. These represent the "human" signal.[44]

In the second step, we presented approximately 200 mTurkers with a set of cue words. For each cue word, we also presented two context words, one human- and one embedding-generated. These words were unlabelled, so that the worker had no information about the source of the context word. Then, we asked the mTurker to indicate the context word that better corresponded with the cue word. Figure F1 provides an example of this "triad task."

Figure F1: Word Embeddings Validation Turing Test: In this triad task, the human respondent observes the cue word "court," and must decide whether "ruling" or "judge" is a more suitable context word.

For each cue word, we calculated the expected probability that the embedding-generated context word was preferred by mTurkers over the human-generated word. Again, following Rodriguez and Spirling (N.d.), we divide this probability by 0.5 to create a metric ranging from 0 to 2. On this scale, a "1" represents human-rater indifference between embedding- and human-generated context words.

[42] We use the 300d, 1.9million dimension vectors available at http://nlp.stanford.edu/data/glove.42B.300d.zip.

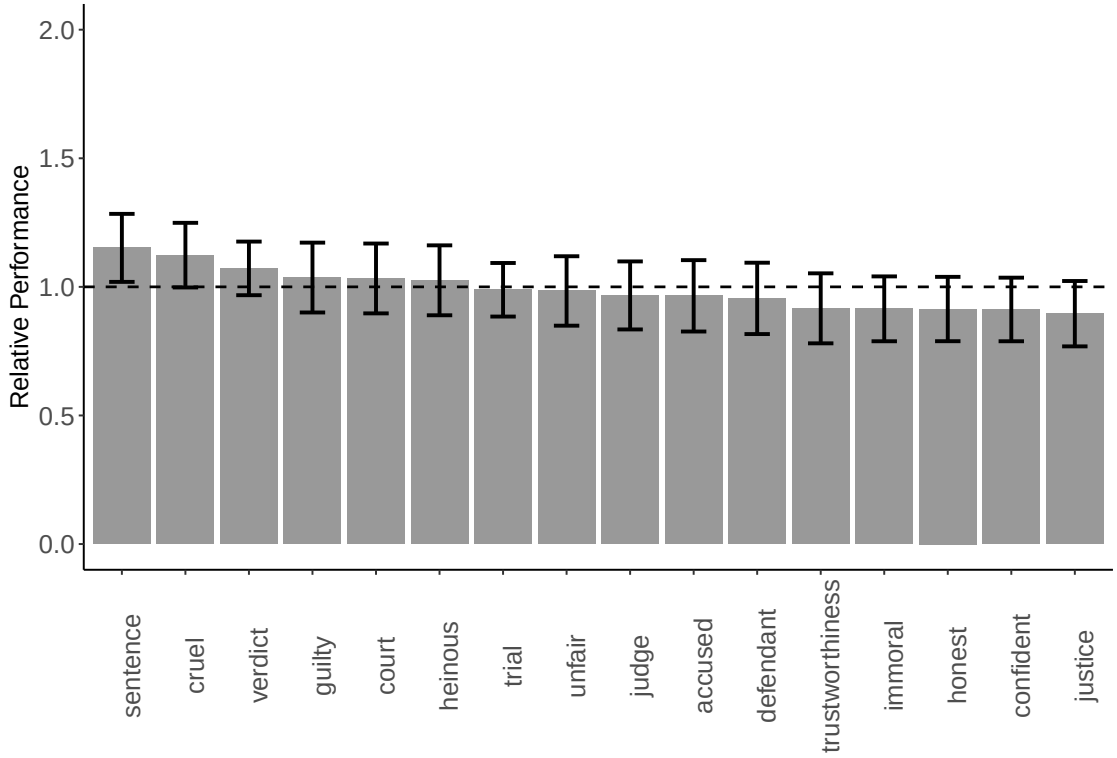
[43] Following the literature, we use the top ten as defined by cosine similarity for the 300d vector.

[44] We thank Pedro Rodriguez for making his Rshiny code available via github for both the generation of human context words and for the Turing test.

Values above one suggest humans prefer the embeddings-generated word; values below one suggest humans prefer the human-generated word.

Figure F2 presents results from the Turing test. Across sixteen words relevant to the legal context in question, we find evidence that the word embeddings compare favorably to human-generated text. For all but one of the 95% bootstrapped confidence intervals contain one, suggesting that mTurkers do not systematically prefer human-generated context words over machine-generated ones.

Figure F2: Word Embeddings Validation Turing Test



Notes: On the y-axis, values equal to one suggest that mTurkers are indifferent between embeddings context words and the human-generated context words on average. Values above one suggest humans prefer the embeddings-generated word; values below one suggest humans prefer the human-generated word. Error bars represent 95% bootstrap confidence intervals.

G Research Ethics

This appendix discusses ethical considerations related to this research. Below, we discuss issues related to the mTurk Turing Test (subsection G.1 and the observational data subsectionG.2), focusing in each subsection on the specific principles considered during the research process. The authors affirm adherence to American Political Science Association (APSA)'s 2020 Principles and Guidance for Human Subjects Research, and have no deviations to report.

G.1 mTurk Turing Test

Principle 11 (Ethical Review) Prior to conducting the mTurk data collection, one of the authors corresponded with the institutional review board regarding whether or not this data collection should undergo full human subjects review. The IRB team advised that as long as mTurk workers are engaged "to provide human validation, or perform a human task that will help validate a statistical model, without 'studying' them [i.e., the mTurkers] in any way (i.e. not collecting any data about them personally, not asking for their perspectives, opinions, beliefs, etc.)," then human subjects review would not be required, though including consent language was recommended. Accordingly, we did not collect any personal data or information on their perspectives, opinions, or beliefs, and we did include consent language.

Principle 5 (Consent) To obtain consent for the human generation of words for the Turing test, we included the following text on the landing page of the data collection website: "This is an academic research project to understand words and their contexts. If you consent to participate in this study, please enter your MTurk ID and press 'Start'."

To obtain consent for the Turing test itself, we included the following text on the landing page of the data collection website: "This is an academic research project to understand how words relate to each other. If you consent to participate in this study, please enter your MTurk ID and press 'Start'."

Additionally, both landing pages contained the following note on confidentiality: "Confidentiality: responses are anonymous, we have no way of linking the data to individual identities."

Principle 6 (Deception) No deception was used during the mTurk Turing Test.

Principle 7 & 8 (Harm and Trauma), Principle 9 (Confidentiality) The Turing Test did not involve any tasks with any potential to inflict harm or trauma on respondents. Furthermore, we did not collect any identifying information for participants; there should be no potential for a breach in confidentiality.

Compensation: Our aim in compensation was to provide a payment equivalent to at least a $10 an hour wage, which would be about 38% higher than U.S. minimum wage. To estimate the time for task completion, one co-author and two undergraduate RAs completed the task.
Word Generation Task: We estimated ex ante that the word generation task would take between 15 and 20 minutes. We paid mTurk workers $3.50 to complete this task, which translated to an hourly wage of between $10.50 and $14. On average, post-completion data show that mTurk respondents took 20 minutes to complete the task, resulting in an average ex post hourly wage equivalent of $10.50.
Turing Test Task: We estimated ex ante that the Turing Test task would take approximately 5 minutes to complete. We paid mTurk workers $1.25 to complete this task, which translated to an hourly wage of between $15 and $18.75 per hour. In practice, this task took approximately 7.77 minutes to complete, resulting in an ex post hourly wage equivalent of $9.65.

G.2 Observational Data

Principle 7 & 8 (Harm and Trauma), Principle 9 (Confidentiality) Although our observational data analysis does not fall under traditional definitions of human subjects research, we nonetheless

briefly discuss the ethical implications of using data on criminal appeals, especially as it pertains to issues of harm, trauma, and confidentiality. Our legal judgements data were accessed on the free, publicly available database at http://kenyalaw.org/caselaw/. This repository is accessible by anyone around the world and was explicitly designed in cooperation with the Kenyan Law Society to make legal decisions in Kenya more transparent. In fact, users can search for cases based on citation, judge name, litigant name, date, and other keywords. For our analysis, we compiled raw text judgements as a spreadsheet, which included the already-public judge and defendant names, along with the requisite indicator variables for case types and court stations. Because our data are derived from a more detailed, structured, and easily accessible public data source, our dataset does not increase existing risks to participants in these legal proceedings. More specifically, the data we generate from these publicly-available records will be less accessible than the easily searchable Kenya Law website (in fact, entering citation information directly into google search redirects you to the Kenya Law website), since the file itself will be contained within a (relatively obscure) political science data archive. Simply put, our data is less accessible, less Google-searchable, and less digestible than the documents that can be found on the public Kenya Law case search website from which our data are derived. We therefore believe that the release of the replication data should not pose any additional harm or trauma (Principles 7 and 8) or breach of confidentiality (Principle 9) beyond the risk imposed by the release of the source data on http://kenyalaw.org/caselaw/.