# Prejudice in Practice

Daniel L. Chen, Jimmy Graham, Manuel Ramos-Maqueda, Shashank Singh*

April 23, 2025

**Abstract**

Why do judges make biased decisions in favor of certain groups? In this article, we examine the role of prejudiced attitudes in driving judicial bias. To do so, we construct a novel dataset of over 150,000 cases from the Kenyan judiciary, use machine learning techniques to analyze written judgements and measure judges' prejudice against women, and leverage the quasi-random assignment of judges to cases to estimate a causal effect of being assigned a prejudiced judge on outcomes for female litigants. We estimate that for every one-standard-deviation increase in a judge's prejudice score, female defendants' probability of winning decreases by approximately 2 percentage points. However, there is no effect for female plaintiffs. This research underscores the influence of prejudice in the public sector and the potential for machine-learning methods to uncover bias.

Word count: 3,987

*Daniel L. Chen (Toulouse School of Economics, daniel.chen@iast.fr); Jimmy Graham (NYU), Manuel Ramos-Maqueda (Oxford), Shashank Singh (University of Chicago)

A large literature has documented that judges often exhibit bias toward specific groups in their decision-making (Shayo and Zussman 2011; Gazal-Ayal and Sulitzeanu-Kenan 2010; Kastellec 2013; Glynn and Sen 2015; Grossman et al. 2016; Yang 2015; Depew, Eren, and Mocan 2017; Arnold, Dobbie, and Yang 2018; Knepper 2018; Sloan 2020; Choi, Harris, and Shen-Bayh 2021). What motivates these biased actions? Scholars have provided various explanations, including a drive to increase the relative standing of one's group (Shayo and Zussman 2011; Chen et al. 2023), empathy for certain groups (Glynn and Sen 2015), and trust in one's group (Choi, Harris, and Shen-Bayh 2021). In this paper, we investigate the role of an under-explored mechanism: prejudiced attitudes. To do so, we construct a text-based measure of prejudice toward women (which we call gender slant) among judges in the Kenyan judiciary and estimate the causal effect of assigning prejudiced judges to cases on outcomes for female litigants.

We build our main dataset by scraping the Kenyan Judiciary's publicly available database for Superior Courts cases over the period 1976-2020 and using machine learning techniques to extract other key variables from the website text.[1] To construct our measure of prejudice toward women for each judge in the dataset, we use word embeddings to estimate the strength of the association between gendered and negative language.

To examine the effect of gender slant of judges on sentencing outcome for litigants, we rely on the quasi-random assignment of cases to Kenyan judges. In Kenya, cases filed to a court are assigned to individual judges based on their existing caseload and the date of filing, which is orthogonal to any other characteristics of the case. Random assignment assures us that any relationship between judges' prejudice and case outcomes is driven by bias rather than other factors, such as self-selection of judges to certain cases. To confirm this, we test for random assignment across gender and we show that male and female defendants and plaintiffs are equally likely to be assigned to a judge regardless of their lexical slant.

We find that slant against women in written judgments is associated with lower win rates

[1] See http://kenyalaw.org/caselaw/.

for female defendants. We estimate that a one-standard-deviation change in the measure of gender slant is associated with about a 2 percentage point decrease in win probability for female defendants. This finding provides evidence that bias in judicial decision-making is driven in part by prejudiced attitudes.

This paper makes several contributions. First, to our knowledge, our paper is the first to evaluate how attitudes toward gender in judicial writings are associated with the corresponding gender bias against female litigants in judicial decisions. By demonstrating that judges with more negative attitudes towards women are more likely to rule against female litigants, we show that judges' attitudes are likely a driving force behind judicial bias. This finding suggests that interventions designed to reduce prejudice (for examples see Paluck et al. (2021)) are a potential means to address bias in judiciaries. It also builds on previous research that has shown that attitudes toward social groups are highly predictive of judgments and choices (Bertrand, Chugh, and Mullainathan 2005), that ideological and biographical characteristics of judges affect their rulings (Boyd and Spriggs. 2009; Glynn and Sen 2015; Kastellec 2013; Sunstein et al. 2007), and that judges displaying gender bias in their writings vote more conservatively in gender-related cases (Ash et al. 2021). We also contribute to the broader literature on the drivers of bias among civil servants (Miller, Kerr, and Reid 1999; Knowles, Persico, and Todd 2001; Plant, Goplen, and Kunstman 2011; Rehavi and Starr 2014; Knox, Lowe, and Mummolo 2020).

Second, our study also demonstrates how machine learning techniques can be leveraged to identify determinants of biased behavior in the judicial context. We show that the use of machine learning in analyzing large-scale legal text allows one to uncover subtle, yet consequential, biases that would be challenging to detect through conventional methods. This quantifiable evidence of prejudice in judicial decisions could inform the targeting of interventions aimed at reducing bias. There is therefore potential for machine learning tools to be integrated into judicial systems to assist in monitoring and mitigating bias in real time, ultimately contributing to more equitable legal processes.

Third, we show where prejudice is activated, specifically, that prejudice is activated against defendants. Defendants in court are frequently associated with wrongdoing, a perception that can trigger deep-seated stereotypes or biases. These biases are often unconsciously linked with ideas of guilt or criminal involvement (Moran and Cutler 1991; Bodenhausen 1990; Philippe and Ouss 2018). The media portrayal of defendants, particularly in crime reporting, tends to emphasize negative stereotypes, influencing public opinion and predisposing people to associate defendants with these negative stereotypes (Dixon 2006). In the courtroom, plaintiffs are generally viewed in a positive light, perceived as seeking justice or redress, while defendants are seen as opposing this pursuit, potentially inviting negative perceptions and the activation of stereotypes. Defendants are often the focal point of negative emotions related to a crime or dispute, such as fear, anger, or loss, which can amplify the activation of stereotypes. Historically, the legal system has exhibited biases, with certain groups more frequently positioned as defendants, resulting in their disproportionate representation and impact within the justice system. This history has the potential to perpetuate specific stereotypes towards defendants (O'Flaherty and Sethi 2022).

# 1 The Kenyan Context

Kenya provides an ideal setting for studying gender-based judicial bias. There is a high degree of gender inequality across a number of dimensions, including representation in the judiciary and a variety of socioeconomic outcomes (IDLO 2020; UNDP 2020).[2] There is also a rich body of written statements from judges that can be used to measure textual slant against women.

The Kenyan judiciary is divided into two main court types: Superior and Subordinate Courts. The vast majority of our data covers the Superior Courts, which include High Courts, which hear both criminal and civil cases and appeals from Subordinate Courts; Environment

[2]According to our data, over the past few decades, female judges have been in the majority for only about 37 percent of cases.

and Land Courts; Employment and Labour Relations Courts; the Court of Appeal, which hears appeals from the High Courts, Environment and Land Courts, and Employment and Labour Relations Courts; and the Supreme Court, which hears appeals from the Court of Appeal and other high-level cases (Kenyan Judiciary 2021).[3]

The Kenyan judiciary does not employ a jury system. This means that judges alone are able to decide the outcomes of cases, which implies that bias among judges can have especially serious consequences. For most cases in most courts, there is only one judge. An exception is in Courts of Appeal, where the majority of cases are composed of multi-judge panels.

# 2 Data

The main data source used in our analysis is the Kenyan Judiciary's publicly available database for court cases.[4] The database includes 159,645 cases over the period of 1976 to 2020. The cases come almost exclusively from the Superior Courts, which comprise the highest level of the court system in Kenya. Kenya Law, an organization within the Kenya Judiciary, began uploading case information in 2006. They upload all cases that are sent to them from the individual courts, and judicial officers in Superior Courts have a mandate to send cases to Kenya Law. For cases prior to 2006, Kenya Law has made (and continues to make) efforts to gather and upload case information.

In order to build our dataset for analysis from this database, we scraped the metadata and full text decision associated with each case. In doing so, we were able to directly extract the following for most cases: the names of plaintiffs, defendants, and judges; the type of case;

[3]According to our data, the Court of Appeals almost exclusively hears civil cases; the Environment and Land Courts are largely split between civil cases and environment and land cases; the Employment and Labour Relations Courts are largely split between labor cases and civil cases; and the High Courts frequently hear a wide range of cases, including civil cases, land and environment cases, labor cases, criminal cases, and others. We have little data on Supreme Court cases, but it appears to hear mostly civil cases. Despite these general trends, the data appears to show that the courts are generally not restricted in the cases they hear, as they all tend to hear a wide range of case types.

[4]See http://kenyalaw.org/caselaw/.

the court in which the case was heard; and the year the judgment was delivered. Our analysis focuses on cases in which there is both a human defendant and human litigant (i.e., cases for which neither litigant is an organization or representative of the state). This approach allows us to study bias toward both defendants and plaintiffs. Since the state is always the prosecutor in criminal cases, criminal cases are not included in our dataset.

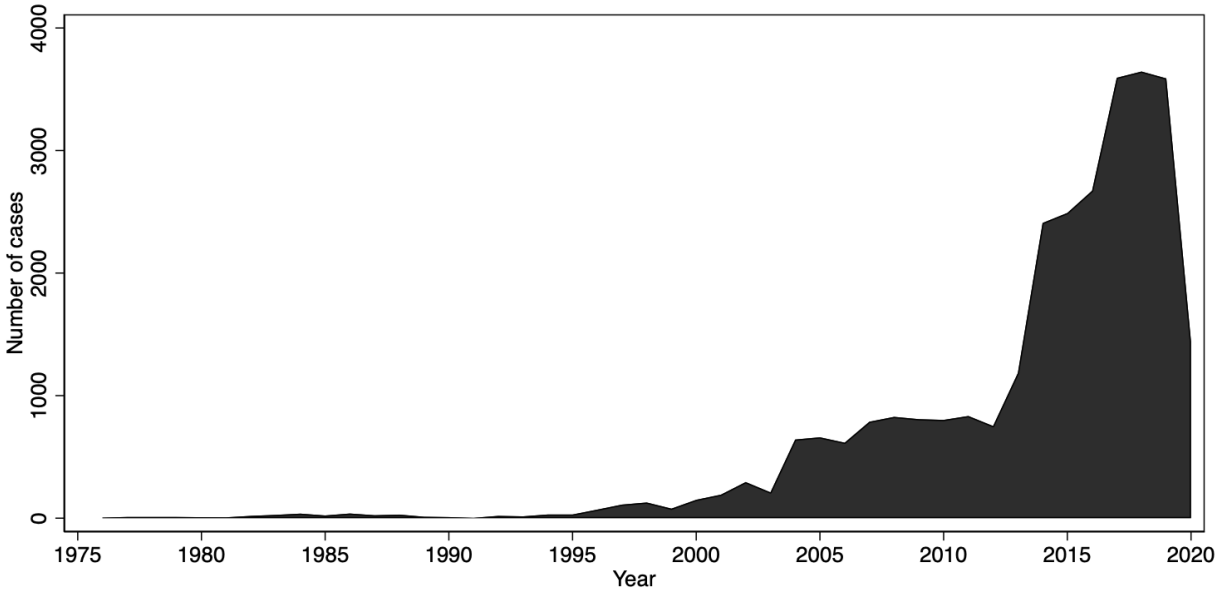
To determine gender and ethnicity and remove non-human cases (i.e., cases with companies, organizations, or the state as litigants), we used the name information scraped from the database. Cases without gender or ethnicity information for judges and either plaintiffs or defendants were dropped. Once gender and ethnicity was assigned to each individual, we could determine the majority genders for the judges, defendants, and plaintiffs for each case.

To measure the gender bias in judges' writing, we follow Elliott Ash, Chen, and Ornaghi (2021) in using a word embedding approach that captures the textual relationship between gendered language and either positive/negative language. This approach allowed us to measure the extent to which judges disproportionately associate women with either negative or stereotypical qualities (e.g.., frivolous, unreasonable, incompetent etc.). The variable resulting from this process is *Median slant*. Positive values indicate greater slant against women. This process is described in greater detail in Appendix A.

We make available the metadata of all 159,645 cases, but focus the analysis on 29,363 cases with litigants who are individuals and have gender or ethnicity data.[5] The data covers 95 courts and 392 judges over the years 1976 to 2020, with an increase in cases over time, as figure 1 shows. Summary statistics of variables in the dataset are presented in Appendix B.

[5]Of the initial 159,645 cases, 33,876 had exclusively human litigants for civil cases. An additional 4,513 cases were dropped because we were unable to determine majority gender or ethnicity for the litigants in the case.

Figure 1: Frequency of cases in the dataset over time




# 3 Empirical strategy

Random assignment is key to our empirical strategy because it assuages the concern that judge slant is correlated with case characteristics that affect outcomes. For example, if judges exhibiting greater gender slant preferred to rule on cases where the male defendants were less likely to be guilty, then we would expect to see indications of gender bias, but the effect would in fact be driven by selection bias. In addition, judges of a certain gender slant may be coincidentally more likely to rule on cases in areas of the country where crime is more or less severe. If these distributions of crime severity are correlated with the gender distributions of defendants and plaintiffs, then we may again falsely perceive gender bias.

## 3.1 Random assignment of cases to judges

To evaluate the existence of gender bias, we exploit the quasi-random assignment of cases to judges. In Kenyan High Courts, cases filed in a court are categorized by court type and sent to the deputy registrar of the relevant court division (family, commercial and

admiralty, labour and employment, constitutional, land and environment, or criminal). The deputy registrar then assigns the case to a judge based on the judge's caseload and calendar, without considering case characteristics. This exogenous assignment is orthogonal to case characteristics such as the gender or ethnicity of the parties. Thus, this system produces as-good-as-random assignment of plaintiffs and defendants to judges, conditional on court division. Our research design therefore allows us to estimate the causal effect of having a more prejudiced judge on litigant outcomes.

To confirm that judge assignment to cases is random in terms of gender majority, we use the following balance test for the analysis sample, for case $i$ filed in court $c$ at time $t$ as:

$$\begin{aligned} judge\_slant_{i,c,t} = \beta_1 def\_maj\_female_{i,c,t} + \\ \beta_2 pla\_maj\_female_{i,c,t} + \Phi_{c,t} + X_{i,c,t,} + \epsilon_{i,c,t} \end{aligned} \tag{1}$$

where $\Phi_{c,t}$ is a court-year fixed effect and $X_{i,c,t}$ is a vector of additional control variables, which may include: binary variables for judge, defendant, and plaintiff plurality ethnicity; judge gender; judge gender interacted with plaintiff and defendant gender, respectively; variables for the numbers of judges, plaintiffs, and defendants; a binary variable indicating whether the case is an appeal; and binary variables indicating the case type. Court-year fixed effects are used to ensure that we are comparing defendants and plaintiffs that are in the same court at the same time. Court-year periods with insufficient variation for regression analysis are dropped from the regressions. For this and all other models, we cluster standard errors at the judge level.

The results of the balance test are shown in Table C1 in the appendix. Column (1) does not include any additional controls. Column (2) includes controls for interactions with judge gender, and Column (3) adds ethnicity and other additional controls (as listed in the table notes). The results indicate that the gender of defendants and plaintiffs is not associated with judge slant. These findings are consistent with the World Bank Doing Business' Index, which asserts that cases are in fact randomly assigned to judges in Kenya (World Bank

2021).

### 3.2 Main specification

To investigate the conditions under which gender bias can be expected, we examine whether judges' slant against women in opinions predicts bias against female defendants and plaintiffs. We model outcome $Y_{i,c,t}$ (where Y=1 corresponds to the defendant winning the case) for case $i$ filed in court $c$ at time $t$ as:

$$\begin{aligned} Y_{i,c,t} = \alpha + \beta_1 judge\_slant_{i,c,t} + \beta_2 def\_maj\_female_{i,c,t} + \\ \beta_3 judge\_slant_{i,c,t} * def\_maj\_female_{i,c,t} + \Phi_{c,t} + X_{i,c,t,} + \epsilon_{i,c,t} \end{aligned} \tag{2}$$

where $judge\_slant$ refers to the mean of the measured lexical slant of the judge group and $def\_maj\_female$ is a binary variable indicating whether defendant group is majority female. The specification used to test the effect of slant towards plaintiffs is identical to (2), except a binary variable for plaintiff majority gender, $pla\_maj\_female$ substitutes $def\_maj\_female$. An alternate specification includes both variables. The main outcomes of interest are the interactions with slant, which indicate whether female defendants and plaintiffs are less likely to win the case if the judge exhibits slant in her/his writing.

## 4 Results

The results of the slant analysis are presented in Table 1. The table provides evidence of a correlation between biased writing and negative outcomes for women. It suggests that, for a 0.05 increase in the judges' slant against women (equivalent to about one standard deviation of the slant measure), female defendants are about 1.6 to 1.8 percentage points less likely to win. The results hold across various specifications.[6]

One potential alternative explanation for our results is that male judges are more preju-

[6]Coefficients for all variables (except fixed effects) are displayed in Appendix D.

diced against women and also rule against women more often than female judges—for reasons unrelated to prejudice against women. However, the fact that the results are robust to the inclusion of judge gender controls (see the final column) undermines this explanation.

Figure 2 presents the predicted win proportions for male and female defendants and various levels of judge slant. These predictions are based on Table 1 column (3). In this case, the figure shows that male defendants are essentially unaffected by a judge's slant. However, female defendants are still less likely to win if judges are more slanted against women in their writing.

Interestingly, this bias seems to be present for female defendants but not female plaintiffs. One possible explanation for this trend is that prejudices toward a particular group may be activated or reinforced when a member of that group is put on trial and accused of wrongdoing. Such an explanation would be consistent with research that shows that certain contextual cues can activate prejudiced attitudes (Lepore and Brown 1997). Defendants are typically in court because they are accused of wrongdoing, which may activate existing stereotypes or biases. People may unconsciously associate certain stereotypes with the idea of being guilty or involved in criminal activity (Moran and Cutler 1991; Bodenhausen 1990; Philippe and Ouss 2018). The portrayal of defendants in media often leans towards negative stereotypes, especially in crime reporting (Dixon 2006). This can influence public perception, making it more likely for stereotypes to be activated when people think of defendants. In the judicial context, plaintiffs are seen as seeking justice or redress, which are generally viewed positively. Defendants, on the other hand, are seen as resisting this process, which can be viewed negatively, activating stereotypes (Bodenhausen 1990). There has been a historical bias in the legal system where certain groups are more likely to be defendants and are disproportionately affected by the justice system. This history can reinforce stereotypes specifically towards defendants. Defendants are often directly connected to the negative emotions surrounding a crime or dispute (fear, anger, loss). These emotions can intensify the activation of stereotypes. In many legal proceedings, the focus is more on the actions

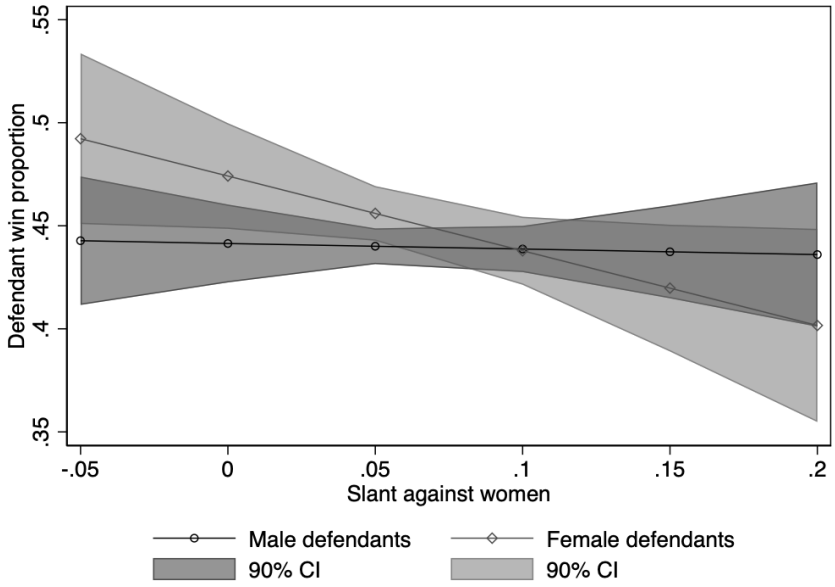
and character of the defendant than on the plaintiff, which can lead to a greater activation of stereotypes (O'Flaherty and Sethi 2022).

Table 1: Main results

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Def. win | Def. win | Def. win | Def. win |
| Pla. maj. female | -0.0338** | -0.0491*** | -0.0347** | -0.0235 |
| | (0.0151) | (0.00975) | (0.0152) | (0.0157) |
| Def. maj. female | 0.0112 | 0.0334** | 0.0327** | 0.0176 |
| | (0.00916) | (0.0136) | (0.0137) | (0.0143) |
| Slant against women | -0.107 | -0.0870 | -0.0269 | -0.0192 |
| | (0.153) | (0.150) | (0.157) | (0.156) |
| Pla. maj. fem. X Slant against women | -0.246 | | -0.226 | -0.232 |
| | (0.180) | | (0.182) | (0.178) |
| Def. maj. fem. X Slant against women | | -0.350** | -0.336* | -0.345** |
| | | (0.169) | (0.170) | (0.172) |
| DV mean | 0.442 | 0.442 | 0.442 | 0.443 |
| Court-year FE | Yes | Yes | Yes | Yes |
| Ethnicity dummies | No | No | No | Yes |
| Other controls | No | No | No | Yes |
| Observations | 15642 | 15642 | 15642 | 15297 |

The regressions test whether defendants/plaintiffs are more likely to lose if they are female and the judge is slanted against females in their writing. The coefficients of interest are on the interaction terms in the last two rows. The measure of slant against women is based on the judges' association of women with negative qualities. All columns are based on a linear regression model. Ethnicity dummies include binary variables indicating whether a given ethnicity is the plurality, one for each ethnicity, for defendants, plaintiffs, and judges. Other controls include case type dummies; a dummy for an appeal case; variables for the numbers of defendants and plaintiffs; an indicator for judge gender; and interactions between judge gender and plaintiff and defendant gender, respectively. To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown. Pla. = plaintiff, def. = defendant, maj. = majority. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Coefficients for all variables (except fixed effects) are displayed in Appendix D.

Figure 2: Predicted defendant win proportions at various levels of judge slant, by defendant gender




Based on table 4, column (3).

# 5 Conclusion

This paper applies machine-learning techniques to uncover an understudied driver of judicial bias: prejudiced attitudes. Employing a novel dataset of over 150,000 cases from the Kenyan judiciary, we build a measure of judicial slant against women that captures the extent to which judges associate women with negative terms in their written judgments. Leveraging the quasi-random assignment of judges to cases, we show that a one-standard-deviation increase in slant reduces female defendants' chances of winning by about 2 percentage points. This bias is not present for female plaintiffs, indicating a specific disadvantage for female defendants.

Our results highlight the impact of prejudice in the public sector, specifically within a judicial setting. While previous research on the specific mechanisms of judicial bias has largely focused on favoritism for certain groups and has even provided evidence against negative attitudes as a mechanism (Shayo and Zussman 2011;Glynn and Sen 2015; Choi, Harris, and

Shen-Bayh 2021; Chen et al. 2023), we show that prejudices can in fact be a significant factor in judicial decision-making. We also demonstrate that machine learning techniques can be used to identify determinants of prejudice in the judiciary. These approaches could be used to target interventions to reduce the impact of prejudice on judges' decisions and create more just judicial systems.

# References


Antoniak, Maria and David Mimno (2018). "Evaluating the stability of embedding-based word similarities". In: *Transactions of the Association for Computational Linguistics* 6, pp. 107–119.

Arnold, David, Will Dobbie, and Crystal Yang (2018). "Racial Bias in Bail Decisions". In: *The Quarterly Journal of Economics* 133.4, pp. 1885–1932.

Ash, Elliot et al. (2021). "Measuring Gender and Religious Bias in the Indian Judiciary". In: *Working paper*.

Ash, Elliott, Daniel Chen, and Arianna Ornaghi (2021). "Gender Attitudes in the Judiciary: Evidence from U.S. Circuit Courts". In: *Working Paper*.

Bertrand, Marianne, Dolly Chugh, and Sendhil Mullainathan (2005). "Implicit discrimination." In: *American Economic Review* 95.2, pp. 94–98.

Bodenhausen, Galen V (1990). "Second-Guessing the Jury: Stereotypic and Hindsight Biases in Perceptions of Court Cases 1". In: *Journal of Applied Social Psychology* 20.13, pp. 1112–1121.

Boyd, Christina and James Spriggs. (2009). "An examination of strategic anticipation of appellate court preferences by federal district court judges". In: *Wash. U. J. L. and Pol'y* 37.

Chen, Daniel et al. (2023). "Re-examining Judicial Bias". In: *Working paper*.

Choi, Danny, Andy Harris, and Fiona Shen-Bayh (2021). "Ethnic Bias in Judicial Decision Making: Evidence from Criminal Appeals in Kenya". In: *American Political Science Review* 116.3, pp. 1067–1080.

Depew, Briggs, Ozkan Eren, and Naci Mocan (2017). "Judges, juveniles, and in-group bias". In: *The Journal of Law and Economics* 60.2, pp. 209–239.

Dixon, Travis L (2006). "Psychological reactions to crime news portrayals of Black criminals: Understanding the moderating roles of prior news viewing and stereotype endorsement". In: *Communication Monographs* 73.2, pp. 162–187.

Gazal-Ayal, Oren and Raanan Sulitzeanu-Kenan (2010). "Let My People Go: Ethnic In-Group Bias in Judicial Decisions–Evidence from a Randomized Natural Experiment". In: *Journal of Empirical Legal Studies* 7.3, pp. 403–428.

Glynn, Adam and Maya Sen (2015). "Identifying judicial empathy: does having daughters cause judges to rule for women's issues?" In: *American Journal of Political Science* 59.1, pp. 37–54.

Grossman, Guy et al. (2016). "Descriptive Representation and Judicial Outcomes in Multiethnic Societies". In: *American Journal of Political Science* 60.1, pp. 44–69.

IDLO (2020). *Women's Professional Participation in Kenya's Justice Sector: Barriers and Pathways*. International Development Law Organization.

Kastellec, Jonathan (2013). "Racial diversity and judicial influence on appellate courts". In: *American Journal of Political Science* 56.1, pp. 167–183.

Kenyan Judiciary (2021). *Courts: Overview*. URL: `https://www.judiciary.go.ke/courts/`.

Knepper, Matthew (2018). "When the shadow is the substance: Judge gender and the outcomes of workplace sex discrimination cases". In: *Journal of Labor Economics* 36.3, pp. 623–664.

Knowles, John, Nicola Persico, and Petra Todd (2001). "Racial bias in motor vehicle searches: Theory and evidence". In: *Journal of Political Economy* 109.1, pp. 203–229.

Knox, Dean, Will Lowe, and Jonathan Mummolo (2020). "Administrative records mask racially biased policing". In: *American Political Science Review* 114.3, pp. 619–637.

Kozlowski, Austin, Matt Taddy, and James Evans (2019). "The geometry of culture: Analyzing the meanings of class through word embeddings". In: *American Sociological Review* 84.5, pp. 905–949.

Lepore, Lorella and Rupert Brown (1997). "Category and stereotype activation: Is prejudice inevitable?" In: *Journal of personality and social psychology* 72.2, p. 275.

Miller, Will, Brinck Kerr, and Margaret Reid (1999). "A national study of gender-based occupational segregation in municipal bureaucracies: Persistence of glass walls?" In: *Public Administration Review*, pp. 218–230.

Moran, Gary and Brian L Cutler (1991). "The Prejudicial Impact of Pretrial Publicity 1". In: *Journal of applied social psychology* 21.5, pp. 345–367.

O'Flaherty, Brendan and Rajiv Sethi (2022). "Stereotypes and the administration of justice". In: *Handbook on Economics of Discrimination and Affirmative Action*. Springer, pp. 1–25.

Paluck, Elizabeth Levy et al. (2021). "Prejudice reduction: Progress and challenges". In: *Annual review of psychology* 72, pp. 533–560.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing*.

Philippe, Arnaud and Aurélie Ouss (2018). ""No hatred or malice, fear or affection": Media and sentencing". In: *Journal of Political Economy* 126.5, pp. 2134–2178.

Plant, E. Ashby, Joanna Goplen, and Jonathan W. Kunstman (2011). "Selective responses to threat: The roles of race and gender in decisions to shoot". In: *Personality and Social Psychology Bulletin* 37.9, pp. 1274–1281.

Rehavi, Marit and Sonja B. Starr (2014). "Racial disparity in federal criminal sentences". In: *Journal of Political Economy* 122.6, pp. 1320–1354.

Shayo, Moses and Asaf Zussman (2011). "Judicial ingroup bias in the shadow of terrorism". In: *The Quarterly Journal of Economics* 126.3, pp. 1447–1484.

Sloan, CarlyWill (2020). "Racial bias by prosecutors: Evidence from random assignment". In: *Working paper*.

Spirling, Arthur and Pedro Rodriguez (2019). "Word embeddings: What works, what doesn't, and how to tell the difference for applied research". In: *Journal of Politics*.

Sunstein, Cass et al. (2007). *Are judges political? An empirical analysis of the federal judiciary*. Brookings Institution Press.

UNDP (2020). *Gender Inequality Index*. URL: http://hdr.undp.org/en/content/gender-inequality-index-gii.

World Bank (2021). *Ease of Doing Business in Kenya*. URL: https://www.doingbusiness.org/en/data/exploreeconomies/kenya#DB_ec.

Yang, Crystal (2015). "Free at Last? Judicial Discretion and Racial Disparities in Federal Sentencing". In: *The Journal of Legal Studies* 44.1, pp. 75–111.

## Competing interests

Competing interests: The authors declare none

## Appendix A: Using word embeddings to determine textual slant

To determine each judge's textual gender slant (i.e., the degree to which each judge exhibits gender bias in their written judgments), we make use of word embeddings, which model the text present in the judgments in the form of low dimensional euclidean space vectors (Pennington, Socher, and Manning 2014). In other words, word embeddings are low dimensional vectors which can accommodate large vocabularies and corpora without increasing dimensionality. The representation resulting from them captures relations between the words. In order to catch semantic similarity amongst words, the positions are assigned to word vectors in the euclidean space, such that the words that appear frequently in the same context have representations close to each other in the space, while words that appear rarely together have representations that are far apart.

To train our word embeddings, we used the GloVe algorithm, described above. The embeddings we trained were then used for identification of cultural dimensions in language (Kozlowski, Taddy, and Evans 2019). That is, we identified a gender dimension by taking the difference between the average normalized vector across a set of male words and the average normalized vector across a set of female words, as such:

$$\vec{male} - \vec{female} = \sum{}_n \vec{maleword}_n / \mid N_{male} \mid + \sum{}_n \vec{femaleword}_n / \mid N_{female} \mid$$

where $N_{male}$ is the number of words used to identify the male dimension. In order to determine the similarity within these dimensions, we used cosine similarity as a measure, defined as follows:

$$sim(\vec{x}, \vec{y}) = cos(\theta) = (\vec{x} \cdot \vec{x}) / (\parallel \vec{x} \parallel \parallel \vec{y} \parallel)$$

where $\vec{x}$ and $\vec{y}$ are non-zero vectors, $\theta$ is the associated angle, and $\parallel \cdot \parallel$ is the 2-norm. Therefore, we can see that words with male (female) connotations are going to be positively

(negatively) correlated with the gender dimension defined by $\vec{male} - \vec{female}$.

These dimensions were then used to construct the gender slant measures. For the first, we aimed to capture the strength of the association between gender attitudes, which identify men more positively and women negatively. Specifically, we used the cosine similarity between the vector representing the gender dimension, defined by $\vec{male} - \vec{female}$, and the vector representing the good-bad dimension, defined by $\vec{good} - \vec{bad}$. We aimed to capture stereotypical attitudes that associate men with "good" and women with "bad" words.

For the $\vec{male} - \vec{female}$ dimension, we used various gender-specific words which were found out to be the five most frequently occurring in our corpus. Words for $\vec{good} - \vec{bad}$ were chosen in a similar fashion. Only five words were chosen for each because, given the relatively small size of the corpus, the inclusion of too many words could result in invalid measures of slant. The word used are displayed in table 2.

Table 2: Words used for each vector dimension

| Vector dimension | Words |
|---|---|
| $\vec{MaleNames}$ | john, joseph, peter, james, david |
| $\vec{FemaleNames}$ | faith, mary, rose, jane, margaret |
| $\vec{Male}$ | his, he, him, mr, himself |
| $\vec{Female}$ | her, she, ms, mrs, herself |
| $\vec{Good}$ | competent, strong, power, serious, professional |
| $\vec{Bad}$ | frivolous, vain, incompetent, unreasonable, incapable |

Each dimensions includes the five most common relevant words in the corpus. Only five words were chosen for each because, given the relatively small size of the corpus, the inclusion of too many words could results in invalid measures of slant.

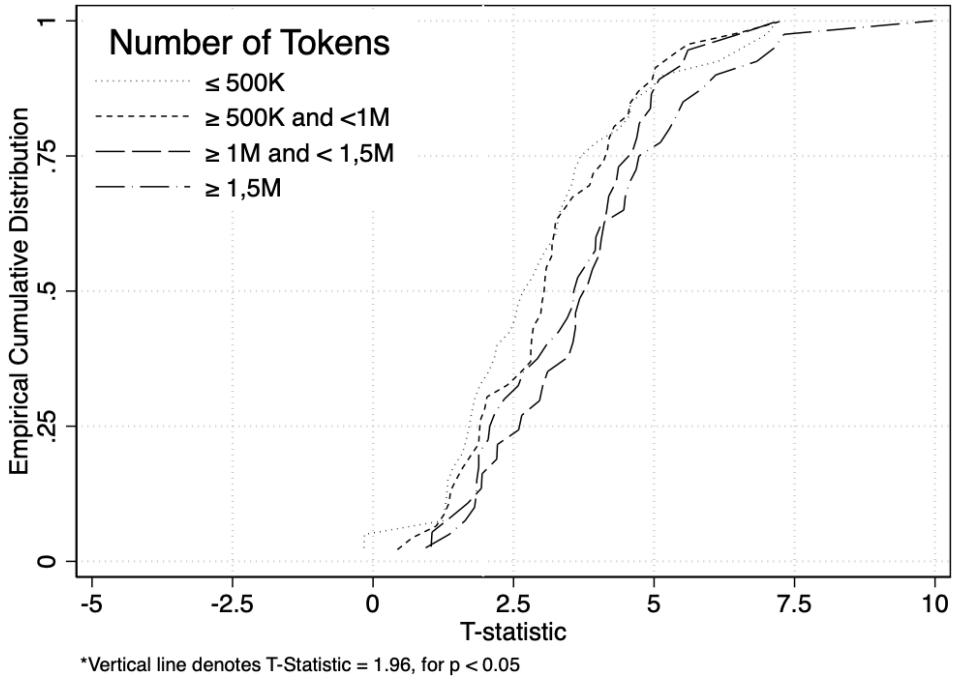
To apply this process to the data, we first preprocessed the entire Kenya Law corpus of judgments by removing punctuations (but retaining hyphenated words). To avoid case sensitivity, we transformed all our words to lower case. We then retained only the most common 50,000 words in all judicial opinions. To obtain judge-specific gender slant measures, we took the set of majority opinions authored by each judge as a separate corpus and trained separate GloVe embeddings on each judge's corpus. To ensure convergence, we trained vectors for 20 iterations with a learning rate of 0.05.

Since each judge might not have a sufficiently large number of tokens, we follow the

approach suggested by Antoniak and Mimno (2018) and train embedding models on 25 bootstrap samples of each judge corpus. Specifically, we consider each sentence written by a judge as a document and then create a corpus by sampling with replacement from all sentences. The number of sentences contained in the bootstrapped sample is the same as the total number of sentences in the original judge corpus. We then calculate our slant measure for all bootstrap samples and assign to each judge the median value of the measure across the samples. Given that embeddings trained on small corpora tend to be sensitive to the inclusion of specific documents, the bootstrap procedure produces more stable results. In addition, bootstrapping ensures stability with respect to the initialization of the word vectors—a potential concern given that GloVe presents a non-convex objective function (Spirling and Rodriguez 2019). The variables resulting from this process is *Median slant*. Positive values indicate greater slant against women.

To validate that the embeddings capture meaningful information about gender, after following the bootstrapping procedure, we compute the cosine similarity between the gender dimension and each of the vectors representing the five most common male and female names for each judge and bootstrap sample. We then regress a dummy for whether the name is male on the median cosine similarity between the vector representing the name and the gender dimension across bootstrap samples, separately for each judge. Figure 3 shows the cumulative distribution of the t-statistics resulting from these regressions for sets of judges with different numbers of tokens. It shows that most t-statistics are significant (and they are never lower than zero). This shows that the gender dimension identified in the embeddings does indeed contain meaningful gender information.

Figure 3: Cumulative distribution of t-statistics from regressions testing the validity of the word embeddings




The vertical line indicates T-stat=1.96, for significance at p<0.05. T-statistics are from regressions between a dummy for whether the name is male on the median cosine similarity between the vector representing the name and the gender dimension across bootstrap samples, separately for each judge.

## Appendix B: Descriptive statistics

Table B1: Summary of main variables

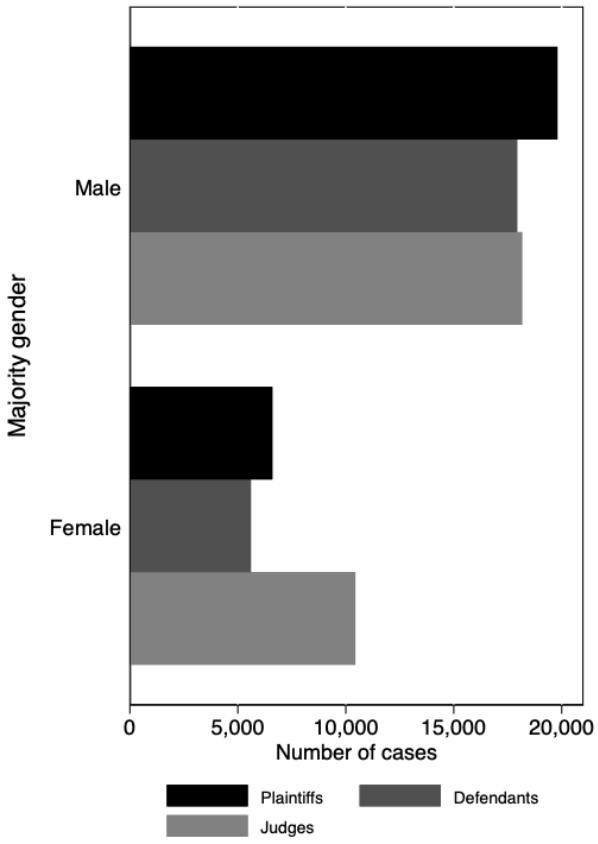
| | count | mean | sd | min | max |
|---|---|---|---|---|---|
| Def. win | 29351 | 0.43 | 0.49 | 0.0 | 1.0 |
| Judge maj. female | 28606 | 0.37 | 0.48 | 0.0 | 1.0 |
| Pla. maj. female | 26401 | 0.25 | 0.43 | 0.0 | 1.0 |
| Def. maj. female | 23541 | 0.24 | 0.43 | 0.0 | 1.0 |
| Judge-plaintiff same ethnicity | 21943 | 0.13 | 0.34 | 0.0 | 1.0 |
| Judge-defendant same ethnicity | 21089 | 0.13 | 0.33 | 0.0 | 1.0 |
| This is an appeal case | 29351 | 0.27 | 0.45 | 0.0 | 1.0 |
| This case is appealed | 29351 | 0.02 | 0.13 | 0.0 | 1.0 |
| Decision is reversed in the appeal | 518 | 0.51 | 0.50 | 0.0 | 1.0 |
| Number of defendants | 29351 | 1.58 | 1.44 | 1.0 | 68.0 |
| Number of plaintiffs | 29351 | 1.31 | 1.13 | 1.0 | 65.0 |
| Number of judges | 29351 | 1.11 | 0.46 | 1.0 | 9.0 |
| Median slant, career v family | 26346 | -0.03 | 0.10 | -0.3 | 0.3 |
| Median slant, good v bad | 22242 | 0.06 | 0.05 | -0.1 | 0.3 |
| Case type: civil | 28490 | 0.46 | 0.50 | 0.0 | 1.0 |
| Case type: tax | 28490 | 0.00 | 0.05 | 0.0 | 1.0 |
| Case type: human rights | 28490 | 0.00 | 0.04 | 0.0 | 1.0 |
| Case type: judicial review | 28490 | 0.00 | 0.03 | 0.0 | 1.0 |
| Case type: divorce | 28490 | 0.00 | 0.04 | 0.0 | 1.0 |
| Case type: election | 28490 | 0.00 | 0.04 | 0.0 | 1.0 |
| Case type: labor relations | 28490 | 0.02 | 0.13 | 0.0 | 1.0 |
| Case type: environment and land | 28490 | 0.32 | 0.47 | 0.0 | 1.0 |
| Case type: family | 28490 | 0.01 | 0.08 | 0.0 | 1.0 |
| Case type: industrial | 28490 | 0.00 | 0.06 | 0.0 | 1.0 |
| Case type: miscellaneous | 28490 | 0.08 | 0.27 | 0.0 | 1.0 |
| Case type: succession | 28490 | 0.10 | 0.29 | 0.0 | 1.0 |
| Number of cases cited in judgement | 29351 | 1.93 | 3.54 | 0.0 | 87.0 |
| Times judgement cited | 29351 | 0.23 | 1.94 | 0.0 | 109.0 |
| Laws cited in judgement | 29351 | 2.20 | 4.11 | 0.0 | 146.0 |
| Words in judgement | 29351 | 1451.78 | 1337.76 | 0.0 | 42980.0 |

Table B2: Frequency of court types in the dataset

| | Frequency |
|---|---|
| Court of appeal | 1659 |
| Employment and labor relations | 1081 |
| Environment and land court | 8619 |
| High court | 17844 |
| Other | 124 |
| Supreme court | 24 |
| Total | 29351 |

Other includes Election Petition in Magistrate Courts, the Judges and Magistrates Vetting Board, Kadhis Courts, and the National Environment Tribunal

Figure B1: Total number of cases, by majority gender and role in the case



## Appendix C: Balance checks

Table C1: Slant balance checks

| | (1) Median slant | (2) Median slant | (3) Median slant |
|---|---|---|---|
| Pla. maj. female | 0.000296 | 0.00148 | 0.00110 |
| | (0.00112) | (0.00140) | (0.00134) |
| Def. maj. female | 0.000234 | 0.000979 | 0.000181 |
| | (0.000766) | (0.000804) | (0.000829) |
| Judge maj. female | | 0.00874 | 0.00851 |
| | | (0.00929) | (0.00923) |
| Pla. maj. fem. X Judge maj. fem | | -0.00304 | -0.00292 |
| | | (0.00247) | (0.00251) |
| Def. maj. fem. X Judge maj. fem | | -0.00220 | -0.00195 |
| | | (0.00194) | (0.00194) |
| Appeal | | | 0.00310 |
| | | | (0.00194) |
| Number of defendants | | | -0.000194 |
| | | | (0.000151) |
| Number of plaintiffs | | | 0.000320 |
| | | | (0.000221) |
| DV mean | 0.0630 | 0.0626 | 0.0626 |
| Court-year FE | Yes | Yes | Yes |
| Judge controls | No | Yes | Yes |
| Ethnicity dummies | No | No | Yes |
| Other controls | No | No | Yes |
| Observations | 15642 | 15297 | 15297 |

Standard errors, in parentheses, are clustered at the judge level. All columns are based on a linear regression model. Judge controls include an indicator for judge gender and interactions between judge gender and plaintiff and defendant gender, respectively. Ethnicity dummies include binary variables indicating whether a given ethnicity is the plurality, one for each ethnicity, for defendants, plaintiffs, and judges. These are essentially ethnicity fixed effects. Other controls include case type fixed effects. To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown. Pla. = plaintiff, def. = defendant, maj. = majority. * p<0.1, ** p<0.05, *** p<0.01

## Appendix D: Full results

Table D1: Main results, all variables displayed

| | (1) Def. win | (2) Def. win | (3) Def. win | (4) Def. win |
|---|---|---|---|---|
| Pla. maj. female | -0.0338** | -0.0491*** | -0.0347** | -0.0235 |
| | (0.0151) | (0.00975) | (0.0152) | (0.0157) |
| Def. maj. female | 0.0112 | 0.0334** | 0.0327** | 0.0176 |
| | (0.00916) | (0.0136) | (0.0137) | (0.0143) |
| Slant against women | -0.107 | -0.0870 | -0.0269 | -0.0192 |
| | (0.153) | (0.150) | (0.157) | (0.156) |
| Pla. maj. fem. X Slant against women | -0.246 | | -0.226 | -0.232 |
| | (0.180) | | (0.182) | (0.178) |
| Def. maj. fem. X Slant against women | | -0.350** | -0.336* | -0.345** |
| | | (0.169) | (0.170) | (0.172) |
| Judge maj. female | | | | -0.0395** |
| | | | | (0.0158) |
| Pla. maj. fem. X Judge maj. fem | | | | 0.00405 |
| | | | | (0.0199) |
| Def. maj. fem. X Judge maj. fem | | | | 0.0505*** |
| | | | | (0.0183) |
| Appeal | | | | 0.0907*** |
| | | | | (0.0154) |
| Number of defendants | | | | 0.00597* |
| | | | | (0.00343) |
| Number of plaintiffs | | | | 0.00346 |
| | | | | (0.00380) |
| DV mean | 0.442 | 0.442 | 0.442 | 0.443 |
| Court-year FE | Yes | Yes | Yes | Yes |
| Ethnicity dummies | No | No | No | Yes |
| Other controls | No | No | No | Yes |
| Observations | 15642 | 15642 | 15642 | 15297 |

The regressions test whether defendants/plaintiffs are more likely to lose if they are female and the judge is slanted against females in their writing. The coefficients of interest are on the interaction terms in the last two rows. The measure of slant against women is based on the judges' association of women with negative qualities. All columns are based on a linear regression model. Ethnicity controls include binary variables indicating whether a given ethnicity is the plurality, one for each ethnicity, for defendants, plaintiffs, and judges. These are essentially ethnicity fixed effects. Other controls include case type fixed effects. To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown. Pla. = plaintiff, def. = defendant, maj. = majority. * p<0.1, ** p<0.05, *** p<0.01