Racial bias in the law is well-established. Studies have demonstrated racial bias in criminal sentencing (Yang, 2015), stop-and-frisk policy (Gelman, Fagan and Kiss, 2007), and the application of police force (Goff et al., 2016). These racial disparities are a function of both explicit biases – conscious expressions of prejudice – and implicit biases – subconscious prejudices “so subtle that people who hold them might not even be aware of them” (Perez, 2016; Eberhardt, Goff and Purdie-Vaughns, 2004, 1196). In this paper, we extend this work by addressing an especially important, and to date unexplored, avenue of implicit racial bias in the law, the words used by judges in their written opinions. Because they provide reasoning for the disposition of current and future cases, “the language of judicial opinions represents the law” (Tiller and Cross, 2006). In order to measure implicit bias in judicial opinions, we leverage an approach to measuring bias from a large corpus (Caliskan, Bryson and Narayanan, 2017). In an initial exploration of over one million judicial opinions from state and federal courts, we find consistent evidence of implicit racial bias. Our work further advances scholarship on racial biases in law by showing that these biases permeate the very language of the law itself, and lays the groundwork for future scholarship, including changes and differences over time, space, and institutional environment.

Implicit Racial Bias

Implicit racial biases are affective evaluations of racial groups that are pre-cognitive, involuntary, and resistant to social desirability pressures (Perez, 2013, 2016). Implicit biases influence a great deal of social and political behavior, with increasing evidence they may “be responsible for many of the continuing racial disparities in society” (Greenwald and Krieger, 2006, 1197). Researchers have employed a number of measures of implicit racial bias (Perez, 2013). Particularly influential is the Implicit Association Test (IAT), in which respondents sort names or words into boxes of “Good”, “Bad”, “Black”, and
“White.” Researchers then measure the response times for sorting, with evidence white Americans take longer to sort when “Black” and “Good” are aligned, and conversely are faster when “Black” and “Bad” are aligned. Using the IAT, researchers have found that implicit racial bias affects vote choices, policy preferences, and jury deliberations (Perez, 2016; Greenwald et al., 2009; Levinson and Young, 2009).

The pervasiveness of implicit racial bias in the mass public begs the question whether similar biases are also evident among political elites. An especially powerful set of elites are judges, whose decisions influence not only the lives of the parties to the case but also the path of legal development and the instantiation of legal interpretations by other political actors. The relation of implicit bias to the judicial exercise of power is particularly concerning given the long-term racial inequalities in the legal system (Kang et al., 2012).

To that end, Rachlinski et al. (2009) administered a variant of the IAT to 133 trial judges, finding strong evidence of a “white preference” among the white judges in their sample. A follow-up experimental vignette, however, did not find evidence implicit racial bias influenced decision outcomes (i.e., guilt/innocence) when race was explicit in the experimental prompt.

Given questions of external validity and small sample size, it remains unclear the extent to which judges express implicit racial biases (Kang et al., 2012), though prior work provides evidence black judges are more likely to have their decisions overturned on appellate review (Sen, 2015). More pointedly, while Rachlinski et al. (2009) and Sen (2015) examine implicit biases in case outcomes, implicit racial bias may lurk in the language of judicial opinions. The stakes of whether implicit racial bias exists in judicial opinions are high. First, the presence of implicit racial bias raises questions about the fairness of the case outcome. Second, each opinion creates precedent – and shapes legal development – through the creation of a legal principle and the explication of the legal reasoning, thereby “help[ing] to structure the outcomes of future disputes” (Hansford and Spriggs, 2006, 3).
This is true both within the courts and in the broader political system as executives, legislators, police, and others adjust their behavior in light of the legal principle and reasoning (Hansford and Spriggs, 2006). Thus, opinions exhibiting racial bias, even if of the implicit variety, may deeply entrench discrimination in law and policy.

We believe implicit racial biases may exist in the texts of judicial opinions. The very nature of implicit bias suggests the difficulty of preventing bias from leaking into one’s actions, particularly when the relevant action is drafting a long and complicated document.¹

At the same time, judicial opinions are also a “least likely” case for the presence of implicit racial bias. Judicial opinions are (most typically) signed public records, featuring statements which have undergone multiple levels of review, and which are drafted by professionals who are particularly sensitive to the potential for bias. With these considerations in mind, we turn to our research design and analysis.

**Word Embeddings**

To assess the presence of implicit racial bias in judicial opinions, we leverage recent advances in computational linguistics on word embeddings, or distributed representations of word meaning (see, e.g., Mikolov, Chen, Corrado and Dean, 2013; Mikolov, Sutskever, Chen, Corrado and Dean, 2013). Word embeddings represent a word as a vector of weights across some $k$ dimensions; each word in a corpus has some distribution of weights across those $k$ dimensions. As intuition, consider three words: “knife”, “spoon”, and “gun.” Both “knife” and “spoon” share semantic meaning on the basis of their use as

---

¹We recognize biases in the texts of judicial opinions could reflect either pre-conscious bias ("implicit associations", as in Greenwald, McGhee and Schwartz (1998)) or intentional strategic rhetoric ("implicit racial appeals” as in Mendelberg (2001)). Given the current (limited) state of knowledge, we begin by assuming biases, if discovered, are pre-conscious. In future work, we plan to use variation in the data to distinguish between implicit biases and intentional racial rhetoric.
utensils, while both “knife” and “gun” share meaning as weapons. Because “knife” and “spoon” appear in similar contexts, they might regularly co-occur with some shared set of words (“food”, for example). Likewise, because “knife” and “gun” appear in similar contexts, they too might regularly co-occur with some other shared set of words (e.g., “murder”). Though the precise procedures vary, approaches to estimating word embeddings use word co-occurrence as an outcome and estimate word-level vectors of weights that best predict those co-occurrences.

Therefore, for each word, we obtain a distribution of weights over a set of dimensions. An attractive feature of this approach is that the vectors associated with two terms that share similar semantic meaning are quite similar. Take a well-known example of working with word embeddings, the following four terms: “king”, “queen”, “man”, and “woman.” One classic example of working with word embeddings is to take the vector of estimated weights across “king”, subtract the vector of weights associated with “man”, and finally add the vector of weights for “woman.” Across implementations, when researchers identify the most similar word vector to this new vector, they find it is the vector of weights associated with “queen.”

Measuring Bias in Corpora

The ability to measure the distance in meaning between two terms provides an avenue to measure bias in text corpora. Consider the word “knife” again. If we were to estimate word embeddings only within a corpus of restaurant reviews, we would expect to find “knife” more closely related to other utensils and food items. On the other hand, if we were to estimate word embeddings using a corpus of judicial opinions, we might reasonably expect to find “knife” more closely related to “gun” or other weapons. Impor-

---

2For a general discussion of word embeddings and evaluation methods, see Vylomova et al. (2016).
tantly, the distributed representation of “knife” would still capture the other dimension of interest – weapons or food, respectively – but it be more similar to the concept that more regularly co-occurs with it in that corpus.

We use this property to measure bias in legal texts. We identify four concepts: African American, white, pleasant, and unpleasant. Each concept is captured by a group of terms drawn from several well-established and well-regarded dictionaries, detailed below. Our approach follows that of Caliskan, Bryson and Narayanan (2017), which borrows from the logic of implicit association tests using racially distinctive names.

Given the sets of terms falling in each of the four categories, the objective is to determine the relative association of the African American category with the pleasant/unpleasant categories in comparison to the association of the white category with the pleasant/unpleasant categories. We do so by first estimating word-specific associations (that is, the association of each African American and white name with each pleasant and unpleasant term). Treating (A,B) as the pleasant and unpleasant categories, we follow Caliskan, Bryson and Narayanan (2017) and calculate word association measures for each name $w$ as:

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})$$

(1)

where $\vec{w}$ represents the the word embedding for name $w$, and $\cos$ represents cosine similarity of the indicated vectors. For both white and African American names, the result is a vector of average association scores between the name $w$ and the pleasant and unpleasant categories. By way of illustration, consider the “knife” example again. The premise is to estimate the difference between the average association of “knife” with food and utensil words and the average association of “knife” with weapons words.

Importantly, by computing the difference in associations within word, the approach has the benefit of minimizing frequency-related complications.
Thus, from the above we have the relative association of individual names within each racial category to the unpleasant and pleasant categories. Interpreting the significance of those relative associations, however, presents a challenge, as we do not know the distribution of associations between African American/white names and pleasant/unpleasant terms under the null hypothesis. To address this, we utilize a series of randomization tests (Lax and Rader, 2010). Randomization tests generate the distribution that we would observe if the null hypothesis was true by randomly changing the data and re-estimating the test statistic, iteratively building a distribution across different random shuffles. Our randomization approach operates as follows. We calculate a relative measure of association between the sets of names and the sets of pleasant/unpleasant terms; following Caliskan, Bryson and Narayanan (2017), we then calculate a word embedding association test statistic as:

\[ s(w, X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \]  

(2)

where \((X,Y)\) indicates white and African American name vectors, respectively. Then, we take the entire, combined set of pleasant and unpleasant terms, and randomly assign those terms into the pleasant and unpleasant categories. That is, we throw out what we know about the terms, and instead assign labels randomly. By then calculating the test statistic with the randomly shuffled data, randomization tests allow us to obtain a test statistic under a situation where the relationship is by definition absent. By repeating this process – here, 10,000 times – of randomly shuffling the assignment of pleasant/unpleasant terms, we obtain a full distribution of test statistics under the null hypothesis. Then, an observed test statistic under the actual division of pleasant/unpleasant words that is at the extreme tail end of the distribution of randomized relationships where there is no effect provides evidence of implicit racial bias.
Data

We acquired all U.S. Supreme Court opinions for the entire history of the Court, all available U.S. Courts of Appeals opinions, and all available state court opinions.\footnote{Texts come from Justia. In the supplementary material, we include an overview of coverage. The number for Pennsylvania is inflated as separate opinions (dissenting, concurring, etc.) are included as separate opinion files. Maryland, New York, and Ohio are excluded, as the opinions posted from those states proved unreliable.} We have 1,134,383 judicial opinions covering 47 states, the U.S. Courts of Appeals, and the U.S. Supreme Court. The entire corpus features 1,630,175 unique tokens, and more than 2.95 billion tokens in all. As points of comparison, recent work in computational linguistics on word embeddings utilized “a 2010 Wikipedia dump with 1 billion tokens; a 2014 Wikipedia dump with 1.6 billion tokens; Gigaword 5 which has 4.3 billion tokens; the combination Gigaword5 + Wikipedia2014, which has 6 billion tokens; and 42 billion tokens of web data, from Common Crawl” (Pennington, Socher and Manning, 2014, 7).

To address concerns that one court or time period may drive the results, we estimate word embedding models across different poolings of the opinions. At the most general level, we estimate models with all opinions across all time periods, then reduce that to just opinions from 1955 to the present, and then reduce further to all opinions from 1969 on. The cutoff years – 1955 and 1969, respectively – are chosen due to their relevance to major events related to civil rights and law (Morris, 1984). In the former case, the landmark case of Brown v. Board of Education was decided in 1954, an event that is widely viewed as inaugurating a new era of judicial activism. In the latter case, 1969 marked the replacement of liberal chief justice Earl Warren by Nixon appointee Warren Burger, inaugurating a more conservative turn in civil rights jurisprudence. Further, we mitigate the influence of individual court levels by estimating word embedding models only for court level across each temporal partition. In all, we estimate and report results from
to twelve separate word embedding models.\(^5\)

**Analysis**

To estimate word embeddings, we utilize GloVe, or Global Vectors (Pennington, Socher and Manning, 2014). This approach first generates a word co-occurrence matrix, with each entry indicating how frequently two words co-occur within some window size (say, within 20 words of one another). The training objective is then to learn word vectors such that the vectors best predict the logarithm of the probability of word co-occurrence (Pennington, Socher and Manning, 2014). We estimate 300-dimensional vectors, training the model using a 50-iteration AdaGrad (Duchi, Hazan and Singer, 2011).\(^6\) We tokenize the text, which involves breaking a document into constituent strings of alphanumeric characters, generally words. We use *tokens* to highlight that different approaches to tokenization may yield different words. We retain only tokens appearing more than 10 times in the corpus, and construct a matrix of co-occurrence counts for tokens co-occurring within a window of 20 tokens.\(^7\)

To capture the African American and white dimensions, we aggregate publicly available lists of stereotypically African American and white names. We utilize five sets of African American and white names, which were all of the available sets we could locate and which were each from prior published work on different forms of racial bias.\(^8\) First, we include the set of names from the Greenwald, McGhee and Schwartz (1998) Implicit Association Test. Second, we use the reduced variant of the Greenwald, McGhee and

---

\(^5\)In the supplemental material, we report results using only the non-aggregated word lists.

\(^6\)AdaGrad stands for adaptive gradient descent. The approach is a popular and widely-utilized gradient-based optimization approach in contexts with sparse data, such as text-as-data approaches.

\(^7\)We utilize a tokenizer that splits based on character, word, line, and sentence boundaries, which is more robust than tokenization based on white space.

\(^8\)The full lists of names and words are available in the supplemental material.
Schwartz (1998) list employed in Caliskan, Bryson and Narayanan (2017). Third, we include an alternative set of names also used in Caliskan, Bryson and Narayanan (2017), itself a derivative of the set of names used in the Bertrand and Mullainathan (2004) study of resume discrimination. Fourth, we use the first names from Butler and Homola (2017), in which the authors validate the use of racially distinctive names in experimental settings. Finally, we employ the set of names from Levitt and Dubner (2005). For pleasant and unpleasant terms, we aggregate four sets of terms. The first two sets are those employed by Caliskan, Bryson and Narayanan (2017) in their Science article, and are selected given their use in prior experimental work on implicit racial bias. One set is from Greenwald, McGhee and Schwartz (1998), and the other set is from Nosek, Banaji and Greenwald (2002). We add a third set from a widely-employed sentiment dictionary (Nielsen, 2011) which features over 1,000 positive and negative terms. Lastly, we use the NRC Emotion Lexicon, which includes over 2,000 positive and negative terms (Mohammad and Turney, 2013).

In Figure 1, we present the results of the randomization tests for statistically significant differences in the association of stereotypically African American names with unpleasant or negative terms in comparison to white names. Each panel presents the results from word embedding models trained across different corpora. Thus, the upper left panel presents the results from a corpus of all opinions across all time periods, while the lower right panel presents the results from a model trained on state court opinions written after 1968. Keep in mind that the distribution is of all test statistics generated from random permutations of the pleasant/unpleasant word sets. Because they are random permutations, they illustrate the distribution under the null effect, and thus are normally distributed around 0. In this set-up, statistically significant racial bias is present in the given corpus if the observed test statistic is in the left tail of the distribution. To illustrate this in the figure below, the vertical dashed grey lines indicate a one-sided p-value of 0.05, while the
Figure 1: Distributions of Test Statistics from 10,000 Iterations These plots provide the distribution of test statistics estimated across random samples of the target characteristic (pleasant/unpleasant). The vertical grey dashed line indicates a one-sided 5% significance test, and the vertical black line indicates the observed test statistic for the indicated opinions.
vertical black lines indicate the actual observed test statistic for the relevant corpus.

Figure 1 makes abundantly clear that racial bias is present in the language of the law. Across all word embedding models based on different court and temporal poolings, the randomization tests provide stark evidence that the observed test statistic is far beyond standard levels of statistical significance. In each model, our observed test statistic is an extreme outlier relative to the distribution under the null hypothesis. Substantively, the conclusion is clear: across judicial institutions and time, judges are much more likely to associate unpleasant terms with African American names, and pleasant terms with white names, in their written opinions. Given the observed differences in association between pleasant and unpleasant terms with African American and white names, and the relative placement of those observed differences across random permutations of the data, we have strong evidence of the same implicit racial bias observed in web corpora present in the language of the law.

Discussion and Conclusion

A rich literature explores the influence of personal beliefs on judicial behavior (e.g., Epstein and Knight, 1998). Likewise, recent work suggests that judges might be susceptible to implicit biases affecting their decision-making (Rachlinski et al., 2009). This paper offers consistent quantitative evidence that implicit racial biases are rife in the language of the law itself. Within and across judicial opinions, African American names are much more frequently associated with unpleasant or negative concepts, while white names are much more frequently associated with pleasant or positive concepts.

These findings have stark implications for our understanding of the “neutrality” of the law from the disposition of particular cases to enforcement of rulings by administrators. Of special concern, our findings suggest that the implicit biases of judges may become en-
trenched in the legal principles and reasoning which guide future decision-making (e.g., Hansford and Spriggs, 2006). This entrenchment may also exacerbate racial inequalities in administrative decision-making and law enforcement.

Our findings also set the stage for important future research. Have implicit biases in judicial opinions changed with shifts in court personnel, alterations in the broader political environment, or the ascendance of the norm of racial equality (Mendelberg, 2001)? Have implicit racial biases in judicial opinions manifested differently across separate regions of the country? Do judges from different racial, ethnic, and gender backgrounds exhibit similar levels of implicit bias? For now, we simply point out that the entrenchment of implicit racial biases in the language of the law imposes yet another obstacle to racial justice.
References


