

Fundamentally Unfair: Analysis of Gendered Discourse in U.S. Capital Trials

Camille Harris¹, Emma Harvey^{2, 3}, Andrea W Wen-Yi³, Sandra Babcock³, Rachel Goldberg³,
Amanda Potts⁴, Allison Koenecke^{2, 3}

¹University of San Francisco

²Cornell Tech

³Cornell University

⁴Cardiff University

charris8@usfca.edu, evh29@cornell.edu, koenecke@cornell.edu

Abstract

Research shows that in capital trials—i.e., cases in which the defendant could face the death penalty if convicted—prosecutors invoke a range of gender stereotypes against women defendants to increase the likelihood of convictions and death sentences. The Supreme Court ruled in *Andrew v. White* (2025) that a prosecutor’s extensive references to a woman defendant’s sex life could be so unduly prejudicial as to render her trial “fundamentally unfair.” In the wake of *Andrew*, evidence of gender bias could be used to advance clemency proceedings or appeals of death sentences. We thus identify an urgent need for systematic methods to measure how gender bias is expressed in the courtroom. As a step towards that goal, we collect a novel corpus of 137 men’s and women’s U.S. capital trial transcripts, including the trials of almost all women currently on death row. We leverage simple NLP techniques to conduct a quantitative analysis of gender-associated key terms used in these trials. We find that, even when conditioning on defendant demographics and crime characteristics, terms associated with *Gender Roles*, *Sexualization*, *Appearance*, and *LGBTQ+ Identity* are used more frequently in the trials of women. Through a supplemental qualitative analysis, we show that the presence of gender-associated key terms can indeed be an indicator of gender stereotyping. Ultimately, we show how simple and interpretable NLP techniques can be used as a tool to identify gender-associated key terms in women’s capital trials. This is a first step towards determining the extent to which gender bias has rendered a trial fundamentally unfair.

1972). Although *Furman* struck down the death penalty, the Court quickly reinstated it in *Gregg v. Georgia* (1976), ruling that the death penalty could be constitutional as long as clear and objective standards were established to guide its use (Gregg 1976). In the years since *Gregg*, however, researchers and advocates have repeatedly found that the death penalty continues to be applied in a way that is not only arbitrary, but also biased. Rather than being tied to characteristics of a crime, the application of the death penalty has historically depended on extrinsic factors such as where the crime occurred, whether defendants could afford to hire their own counsel, and the race of victims and defendants (Bright 1994; Flanders 2019; Pierce, Radelet, and Sharp 2017). Despite this widespread evidence of bias, the Supreme Court ruled in *McCleskey v. Kemp* (1987) that aggregate statistics about the application of the death penalty across demographics (“disproportionate impact”) could not serve as grounds to appeal a death sentence; rather, appeals must be based on a “discriminatory purpose” (McCleskey 1987). As a result, individuals seeking to appeal death sentences on discrimination grounds must demonstrate that discriminatory intent affected their specific case, rather than relying on statistical evidence of systemic disparities in capital sentencing.

While much prior scholarship on bias in capital trials has focused on racial bias (Baldus, Pulaski, and Woodworth 1983), a growing body of work explores gender bias—an issue that is increasingly pressing as a higher proportion of women are incarcerated and given extreme sentences (Jones 2022). Gender bias is often perpetuated through the use of gender stereotyping language in the courtroom. This language can lead courts and juries to substitute cultural expectations about gender for individualized assessments of culpability and credibility, contributing to systematic discrimination against women in the U.S. criminal punishment system (Atwell 2007; Jenkins and Davidson 1990; Jones 2022; Mogul 2005; Potts and Weare 2018).² Providing some relief, the Supreme Court’s 2025 decision in *Andrew v. White* held for the first time that unduly prejudicial gender stereotyping can render capital trials “fundamen-

Appendix — <https://tinyurl.com/gender-aies2026>

1 Introduction

In the U.S. criminal punishment system,¹ the death penalty has long been seen as arbitrary. The U.S. Supreme Court ruled in *Furman v. Georgia* (1972) that the death penalty is “cruel and unusual in the same way that being struck by lightning is cruel and unusual,” that is, applied to a “capriciously selected random handful” of defendants (Furman

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹The term *criminal punishment system* comes from abolitionist work and is increasingly used in legal scholarship to center the system’s fundamentally punitive nature (Foran, Kaba, and Naples-Mitchell 2021; Levin 2023).

²For example, Jenkins and Davidson (1990) found that women accused of killing abusive partners were systematically denied access to self-defense narratives that courts extended to men, and ultimately received harsher sentences as a result.

tally unfair” (Andrew 2025).³ As a result, existence of gender-biased discourse during a woman’s capital trial is now relevant to establish that her due process rights were violated. Already, legal teams representing women on death row in the U.S. are raising new legal arguments based on *Andrew v. White* in clemency proceedings, motions for resentencing, habeas petitions, and pre-trial motions. This underscores the urgent need for systematic methods to identify how gender bias manifests in courtroom proceedings. As a step towards that goal, in this work, we conduct a mixed-methods study leveraging simple NLP methods to analyze a novel corpus of 137 capital trial transcripts—including the trials of almost every woman currently on death row in the U.S.—to answer the following research question: **is there a gap in the frequency of gender-associated terms across women and men defendants in U.S. capital trials, when conditioned on demographic and case-specific features?**

Our work covers several related concepts. *Gender bias* (or *gender discrimination*) is the broad phenomenon of discrimination on the basis of gender. We study one aspect of gender bias, *gender stereotyping*, which we define as an invocation of reductive but commonly held beliefs about a person or group based on their gender and on how their gender is positioned in society (Ammons 1995; Cignarella, Giachanou, and Lefever 2025; Fiske 1998). To measure gender stereotyping, we focus on *gendered discourse*, which is speech that includes *gender-associated key terms*, or terms whose presence could indicate the presence of stereotyping. To do this, we identify stereotyping themes informed by academic literature (Babcock 2024; Potts and Weare 2018) and an analysis of the language used in our corpus. For each theme, we identify associated key terms. We calculate and compare the relative frequency of key term use in the trials of women and men defendants across multiple subsets of our corpus to understand how key terms are used in the trials of women and men who have similar demographics, are convicted of committing similar crimes, and have similar relationships to the victims of the crimes.

We find that, even when accounting for defendant and crime similarity, terms associated with *Gender Roles*, *Sexualization*, *Appearance*, and *LGBTQ+ Identity* are consistently used more frequently in the trials of women. Terms associated with *Characterization*, *Emotions*, *Parenting*, *Intimate Relationships*, *Poverty & Spending*, and *Religion* are almost always used more frequently in the trials of women. Importantly, while this differential prevalence does not necessarily imply that gender stereotyping in general more commonly affects women than men, it provides evidence that women and men defendants are spoken about differently, on average. Using simple methods to identify these patterns at scale is a first step towards showing that a trial may have been “fundamentally unfair.” As a next step,

we conduct a supplemental qualitative analysis in which we manually annotate passages identified as containing gendered discourse by our quantitative analysis, finding that the presence of gender-associated key terms can indeed be an indicator of gender stereotyping. The goal of our work is to provide a template for a simple, repeatable, and interpretable approach to surfacing empirical evidence of gender bias that might be useful in removing defendants from death row.

2 Background and Related Work

The U.S. is among the world’s top executioners, having carried out 1,654 executions between 1976 and 2025 (CCDPW 2018; DPI 2025b). Although women make up only about 2% of people on death row in the U.S. (Babcock 2024), a growing proportion of women have been incarcerated and given extreme sentences in recent years (Jones 2022).⁴

Gender Bias Against Women Capital Defendants. Prior legal scholarship has explored gender bias faced by women on death row, finding that prosecutors, in an effort to convict women defendants, may invoke language that leverages women’s sexuality; relationship to motherhood; or overall lack of adherence to traditional gender norms that expect women to be passive, nurturing, and submissive (Atwell 2007; Babcock 2024; Basow 1992; Jenkins and Davidson 1990; Mogul 2005; Potts and Weare 2018). For example, due to gendered expectations that women possess a maternal instinct, parenting is often invoked to speak negatively about women’s character (Basow 1992)—the “bad mother” narrative is a common framing applied to women tried for murder (Easteal et al. 2015). Similarly, prosecutors may seek to leverage gendered expectations that “good women” are not sexually promiscuous. For example, Greenfield and Babcock (2026) found evidence that prosecutors used women’s sexuality, including their choice of undergarments, their number of prior sexual partners, and their adulterous relationships, to negatively characterize capital trial defendants. Gender stereotypes can be invoked in uniquely harmful ways for women at the intersection of multiple identities that face discrimination in the U.S., including LGBTQ+ women and women of color (Sutton et al. 2021). For instance, Mogul (2005) has found that women’s lesbian identity can be unnecessarily highlighted by prosecutors as evidence of their guilt, while Farr (2000) has found that prosecutors in capital trials describe lesbian defendants as masculine and “man-hating.” Negative stereotypes about racial identity are used to motivate the convictions of Black, Latina, and Native women, while stereotypes about national origin are directed at immigrant women (Rosenthal and Lobel 2016; Slakoff 2020). While our work focuses on gender bias in the U.S., much research also exists exploring gender bias in the court systems of other countries including but

³During Brenda Andrew’s trial for the murder of her husband, prosecutors elicited testimony calling Andrew a “hoochie” and describing her attire as “short skirt, low-cut tops, just sexy outfits” before holding up Andrew’s thong and asking whether it looked like something a “grieving widow” would wear (Andrew 2025). The state has since conceded that much of the evidence it introduced about Andrew’s sex life was not relevant (Andrew 2025).

⁴Men continue to make up the vast majority of individuals on death row in the U.S., and the intent of our work is not to imply that the death penalty is “worse” for women defendants. We make a comparison between women and men to show that gendered discourse may undermine the fairness of women’s trials. We encourage others to analyze the ways that language deployed in men’s cases may render their trials unfair.

not limited to Brazil (Benatti et al. 2024), Portugal (Pinto et al. 2020), and the U.K. (Potts and Weare 2018).

Text Analysis to Identify Gender Bias. Our study builds on prior work that uses both computational NLP methods (Bosco et al. 2023; Cignarella et al. 2024; Deshpande et al. 2023; Fraser, Nejadgholi, and Kiritchenko 2021; Fraser, Kiritchenko, and Nejadgholi 2024) and qualitative analysis (Lucy et al. 2020) to explore bias in text data. In particular, we build on NLP work that has focused on using computational methods to identify various forms of gender stereotyping. Both Bolukbasi et al. (2016) and Caliskan, Bryson, and Narayanan (2017) propose methods for measuring gender stereotyping in text embeddings by comparing the distances between embeddings of gendered words (e.g., “girl” and “boy”) and embeddings of words with known stereotypical or affective valences. Other approaches focus on larger chunks of text. For example, some computational methods attempt to score the language polarity (e.g., sentiment or regard) of text written about a particular social group in order to identify the presence of stereotyping (Sheng et al. 2019). Computational methods may also map to social science theories about how stereotypes are expressed. For example, Fraser, Kiritchenko, and Nejadgholi (2022) score the extent to which text written about a social group expresses perceptions of warmth and competence, which are two axes of stereotyping according to the stereotype content model (Cuddy, Fiske, and Glick 2008; Fiske et al. 2002). Similarly, Mendelsohn, Tsvetkov, and Jurafsky (2020) map psychological components of dehumanization (e.g., referring to members of a social group as non-human) to linguistic features in order to computationally identify instances of dehumanization. Finally, some scholars have used simpler methods that require less NLP expertise in order to identify stereotyping. In particular, Fast, Vachovsky, and Bernstein (2016) count gender-associated key terms made in reference to individuals of different genders in order to quantify gender stereotyping. Because this approach is effective but simple enough to be implemented and understood by individuals without NLP expertise (i.e., lawyers who may seek to identify stereotyping in their clients’ trials), this is the approach we adopt in our work.

While much work on stereotyping starts with text that is already written about a particular social group, identifying relevant text in the first place is also an NLP challenge. A common approach, and one that we adopt in our work, is to use term-level search to identify potentially salient text and then conduct further analysis of that text (e.g., Locatelli, Damo, and Nozza 2023). Although NLP scholars may use computational approaches, such as sentiment prediction, for that analysis, in our mixed-methods study we conduct a manual thematic analysis of relevant passages instead. In this, we are in alignment with prior qualitative work on gender stereotyping (Aley and Thomas 2021; Andina-Diaz et al. 2023; LeBeau, Carr, and Hart 2020), including gender stereotyping in legal contexts (Atwell 2007; Babcock 2024; Eastal et al. 2015; Jenkins and Davidson 1990; Mogul 2005; Potts and Weare 2018; Rosenthal and Lobel 2016). Most similar to our work, Wen-Yi et al. (2024) use

mixed methods to examine gendered narratives, including stereotypes, in a small sample of women’s capital trial transcripts. They leverage qualitative critical discourse analysis and fine-tune LegalBERT to classify gendered stereotypes in text. We extend this work to a larger set of capital trial transcripts, covering over 100,000 pages of text and including the trials of almost all women currently on death row. In doing so, we show how even simple computational methods can be effectively used to identify gendered discourse in U.S. capital trial transcripts.

3 Methods

Data. The legal scholars co-authoring this study collected over 100 U.S. capital trial transcripts through personal outreach, including the transcripts of all people sentenced to death between 1990 and 2022 who were legally recognized as women at their trials. This resulted in our corpus of U.S. capital trial transcripts of 48 women and 89 men defendants.⁵ Each transcript contains the *guilt phase* (i.e., the trial) and the *penalty phase* for each defendant. For each trial except one,⁶ transcripts include opening statements, witness testimony and cross-examination, and closing statements by both the prosecutor and the defense. On average, transcripts in our corpus contain over 450,000 words each (or roughly 1,000 single-spaced pages), making manual analysis at scale difficult. Furthermore, there is no standardized trial transcript format. Many transcripts are decades old, and are available only as PDF scans of typed or even typewritten pages. Before analyzing transcripts, we manually merged transcripts from different phases of a defendant’s trial. We then used Allen AI’s olmOCR 2,⁷ a state-of-the-art open source optical character recognition (OCR) model, to convert transcript PDFs to plain text files (Poznanski, Soldaini, and Lo 2025). Due to the long length and relatively small number of transcripts in our corpus, we chunked each transcript into 500-word documents that form the basis of our analysis.

To better contextualize transcript data and to facilitate richer comparisons of language across defendants, we also collected relevant metadata. For each trial, we identified aggravating factors of the crime (the crime itself is in all cases murder) as described in trial transcripts. Aggravating factors can relate to the victims of the crime (e.g., child

⁵The 48 women’s transcripts include the transcripts of every defendant recognized as a woman at the time of their trial who is currently (as of July 2026) on death row, including 40 cisgender women and 1 transgender man currently on death row. They also include 7 cisgender women who were originally sentenced to death but have since died or had their sentences reduced (DPI 2025a). We were not able to collect the transcripts of transgender women. The 89 men’s transcripts represent about 4% of the approximately 2,000 men currently on death row in the U.S. (The Legal Defense Fund 2025). Because the transcripts were collected through personal outreach, our corpus represents a convenience sample of men’s transcripts and should not necessarily be considered representative of the ways in which all men are spoken about in capital trials.

⁶One woman’s trial transcript was complete except for the penalty phase testimony and closing arguments.

⁷<https://olmocr.allenai.org/>

victim, multiple victims), motives of the crime (e.g., premeditated, retaliation), additional crimes (e.g., rape, kidnapping), the manner of the crime (e.g., crimes that are heinous, atrocious, or cruel⁸), or the weapon used (e.g., gun).⁹ We also collected data on the relationship between the defendant and the victim(s). Finally, using state records and data from the Death Penalty Information Center (DPI), we collected the racial and ethnic demographics of both the defendant and the victim(s).¹⁰

Matched Data Samples. We evaluated patterns of gender-associated key terms within sub-samples of our corpus constructed using the metadata described above. We did this to understand how language used across women's and men's trials differs after controlling for factors such as defendant demographics and characteristics of the crime. Our match procedure (described in more detail below) was to select all women with a particular demographic quality or case characteristic and randomly sample from the set of men with the same features until we reached a balanced sample of women and men. In cases where the set of defendants with a particular feature was very small, we did not strictly enforce balance between women and men in order to include as many transcripts as possible in our analysis. We focus on the relative frequency of gender-associated key terms throughout (i.e., frequency controlling for number of documents) to avoid drawing conclusions based on the number of defendants in each sample.

Crime Match (40 women, 40 men). We examined the subset of women and men convicted of committing similar crimes, which we defined as crimes with the same set of aggravating factors. 31 unique aggravating factors are represented in our corpus. To simplify matching, we condensed similar aggravating factors into an aggregated "crime type" and matched defendants with the same crime types using exact matching (in other words, all of the members of a given crime-matched set had all of the same aggregated crime types).

Parent Defendant Match (41 women, 46 men). While prior work has shown that language of motherhood is often invoked for women on death row, even if it has little to do with their case (Potts and Weare 2018), minimal research has been done to understand if this invocation of parental language is unique to women. Thus we constructed a sample consisting of all defendants who are parents.

Child Victim Match (14 women, 9 men). Similarly, we constructed a sample consisting of all defendants who are accused of murdering their own child or children.

Intimate Partner Victim Match (9 women, 6 men). We matched women and men defendants who were in a current

or prior intimate relationship with the victim.

Defendant Race Match (42 women, 42 men). Prior work has shown how people of color are negatively stereotyped within the U.S. criminal punishment system (Pierce, Radelet, and Sharp 2017), but minimal work focuses explicitly on women of color. We therefore constructed a sample of defendants matched by racial group, featuring 28 white defendants, 8 Black defendants, and 6 Latine defendants of each gender (we excluded Native American and Asian individuals from the match due to insufficient sample size).

Defendant Race and Crime Match (28 women, 28 men). We combined the crime match with the defendant race match, resulting in a sample of 17 white defendants, 8 Black defendants, and 3 Latine defendants of each gender.

Theme and Key Term Selection. To conduct our analysis, we began by identifying themes that may be indicative of gender stereotyping. We derived these themes through an iterative, mixed-methods process leveraging both deductive methods, informed by prior legal scholarship, and inductive methods, informed by computational NLP. First, we developed an initial set of themes corresponding to ways in which stereotyping language has been shown to be used against women during capital trials in prior literature (Babcock 2024; Babcock, Greenfield, and Adamson 2024; Potts and Weare 2018). Additionally, the authors who are legal scholars manually annotated a subset of women's transcripts for the presence of gender stereotyping or gendered discourse. The annotations were often, but not always, mappable to one of the literature-derived themes. We used the unmapped passages to identify additional themes not captured by our initial list. Finally, we used BERTopic (Grootendorst 2022) to identify clusters of words that might represent additional themes.¹¹ All authors reviewed the resulting candidate set of themes and aligned on 16 themes: *Gender Roles, Sexualization, Appearance, Characterization, Emotions, Parenting, Intimate Relationships, LGBTQ+ Identity, Mental Health & Disability, Drugs & Addiction, Poverty & Spending, Racial Tropes, Immigration, Religion, Metaphor, and Violence*. Themes are summarized in detail in Table 1.

For each theme, we identified a set of gender-associated key terms such that the presence of a key term could indicate the presence of the theme, again beginning with the related literature. We then used computational methods to expand the set of key terms. First, we trained a word2vec (Mikolov et al. 2013) model on the corpus to identify words present in the corpus that were most similar, according to cosine similarity, to the initial key terms.¹² Second, we applied guided BERTopic (Grootendorst 2022) to the full set of transcripts, using all of a theme's initial key terms as the guiding terms. Finally, using the manually annotated pas-

⁸Heinous, atrocious, or cruel (HAC) is a legal term that implies torture or serious physical abuse (Bjerregaard et al. 2026).

⁹Aggravating factors are summarized in OSF Appendix Table A1.

¹⁰A list of referenced state and other relevant records is available in OSF Appendix Table A2. DPI data derives from: <https://deathpenaltyinfo.org/death-row/overview/death-row-usa>. We recognize that race is a social construct and that our categorizations may not reflect defendants' and victims' complete identities.

¹¹Prior to training BERTopic, transcripts were lowercased, non-alpha characters were removed, and stopwords and other common words related to the trial context (e.g., 'judge,' 'jury') were removed. Transcripts were chunked into 500-word documents.

¹²Prior to training word2vec, transcripts were lowercased, non-alpha characters were removed, and stopwords and other common words were removed. Transcripts were chunked into 20-word sentences. We used a vector size of 300 and a window of 5.

Theme	Justification	Example Key Terms
<i>Gender Roles</i>	Language that frames women as unfeminine, masculine, or outside expected gender roles may be stigmatizing (Carll 2003; Eastal et al. 2015).	aggressive, assertive, callous, cold
<i>Sexualization</i>	References to sexual history or conduct may be used to hypersexualize women (LeBeau, Carr, and Hart 2020; Rosenthal and Lobel 2016; Babcock, Greenfield, and Adamson 2024; Greenfield and Babcock 2026).	affair, flirt, seduce, sexual
<i>Appearance</i>	References to appearance may be used to sexualize or criticize femininity and moral character (Greenfield and Babcock 2026).	bra, dress, jeans, nails
<i>Characterization</i>	Language about personality and disposition may be leveraged to portray women as manipulative or morally deficient, and in particular may feed into the “women as manipulative schemers” trope (Babcock 2024).	calculated, dishonest, manipulate, selfish
<i>Emotions</i>	Emotional expression, or its absence, may be framed as evidence of guilt, manipulation, or moral defect (Eastal et al. 2015).	breakdown, calm, hysteria, irrational
<i>Parenting</i>	Parental status may be used to judge women against expectations of motherhood and caregiving (Potts and Weare 2018).	caretaker, cps, discipline, neglect
<i>Intimate Relationships</i>	Prosecutors frequently invoke women’s intimate relationships in an effort to portray them as promiscuous or discredit claims of gender-based violence (GBV) (Greenfield and Babcock 2026). We use this theme for more subtle language than outright <i>Sexualization</i> that refers to dynamics within intimate partner relationships.	abusive, domestic, intimate, protect
<i>LGBTQ+ Identity</i>	References to queer identity may be used to portray defendants as deviant, aggressive, or defeminized (Mogul 2005; Eastal et al. 2015).	bisexual, butch, gay, lesbian
<i>Mental Health & Disability</i>	Mental illness and disability may be minimized or weaponized by prosecutors (Baumgartner et al. 2017; Hunt 2023; Farr 2022).	brain, crazy, functioning, iq
<i>Drugs & Addiction</i>	Drug use and addiction is heavily stigmatized in the U.S., with compounding effects for Black defendants in particular (Avery 2019).	crack, dealer, meth, wine
<i>Poverty & Spending</i>	People in poverty make up the vast majority of those on death row in the U.S., and the language of poverty may be used to imply guilt, instability, or moral failure (Babcock, Greenfield, and Adamson 2024).	greedy, lazy, motel, welfare
<i>Racial Tropes</i>	Racial tropes may be evoked against defendants of color. Racial tropes against Black women emphasize “aggression” and “anger” which can be leveraged to support claims of violent behavior or downplay experiences of GBV (Rosenthal and Lobel 2016; Ammons 1995). Indigenous and Latine defendants also face documented discrimination (Franklin 2013). We note that many racial tropes are male-associated in addition to being racialized.	animal, barrio, gang, thug
<i>Immigration</i>	Xenophobic language may be leveraged against immigrant defendants.	accent, alien, culture, illegal
<i>Religion</i>	Religious language may be used to frame women as sinners or morally deviant, especially in cases where prosecutors highlight extramarital relationships (Greenfield and Babcock 2026).	bible, cain, pray, sin
<i>Metaphor</i>	Similar to religious language, metaphors may be used to moralize behavior and intensify condemnation.	angel, fallen, fruit, nightmares
<i>Violence</i>	Nearly all women on death row in the U.S. have experienced GBV in their lifetime, but prior work finds that defense attorneys often fail to adequately discuss how GBV influences the lives of defendants (Babcock and Greenfield 2023). Instead, violence is often invoked to negatively characterize the defendant as dangerous or blameworthy (Franklin 2013; Rosenthal and Lobel 2016; Eastal et al. 2015; Avery 2019; Babcock, Greenfield, and Adamson 2024).	belt, bruise, fist, stab

Table 1: Gender stereotyping themes and example gender-associated key terms analyzed in this work.

sages, we used c-TF-IDF (see below) to identify terms that occurred disproportionately often within passages annotated as containing a particular theme. We then compiled and reviewed the full list of 440 key terms. Prior to analysis, we eliminated terms that did not occur in at least two women's and men's transcripts, and terms for which more than 50% of their use was in a single transcript. The full list of terms is available in OSF Appendix Table A3.

Quantitative Key Term Analysis. To understand the prevalence of key terms within the corpus, we calculated class Term Frequency-Inverse Document Frequency (c-TF-IDF) of each key term. c-TF-IDF is an adaptation of standard TF-IDF, which measures how frequently a word is used in a given document inversely weighted by how frequently the word is used in the corpus as a whole. The class modification treats all documents within a given class (in our case, defendant gender) as a unit. Thus, c-TF-IDF assigns higher scores to words that occur frequently in the transcripts of defendants of a particular gender and rarely in the transcripts of defendants of the other gender.

We also calculated metrics at the theme level. We did this by mapping all key terms to their associated theme (note that a term may be associated with more than one theme) and treating all terms associated with a theme as a single conceptual 'term' (following similar logic to other adaptations of TF-IDF, such as Synset Frequency-Inverse Document Frequency (Capelle et al. 2012)). We then calculated c-TF-IDF at the theme level.

To evaluate statistical significance of results, we created a 95% confidence interval for both term- and theme-level metrics using bootstrapping, repeating calculations 1000 times and taking the 2.5th and 97.5th percentiles of results as the upper and lower bounds. To account for the fact that we conducted multiple comparisons, we controlled the false discovery rate using the Benjamini-Hochberg (BH) procedure.¹³

Qualitative Passage Analysis. The presence of a key term does not itself establish the presence of stereotyping, and the presence of stereotyping does not itself establish a "fundamentally unfair" trial. Instead, key terms function as indicators of evidence that legal scholars have identified as relevant to claims of gender bias in capital sentencing. To determine whether the presence of key terms corresponds to stereotyping, we also conducted a small-scale, supplemental qualitative analysis of passages from our transcript data. We randomly sampled passages containing gender-associated key terms and performed deductive thematic content analysis on them (Braun and Clarke 2006). At a high level, this was intended to examine:

1. Whether terms are used in the context of the theme (e.g., "I saw him punch her" is an example of a *Violence*-

¹³We note that the use of chunked documents can overstate the significance of differences due to disrupting the assumption that data are independent and identically distributed (IID). We nevertheless chose this approach because of relatively small sample size of transcripts makes comparison infeasible otherwise. However, we interpret statistical significance cautiously and focus primarily on practical significance, or meaningfully large differences between point estimates, to account for this.

related use of the word *punch*, while "she served us punch" is not);

2. Whether terms are used in a way that is relevant to the facts of the case; and
3. Whether terms are being used in a way that points to an aggravating or mitigating factor (e.g., "she suffered from battered woman syndrome," is an attempt to present *Mental Health & Disability* as a mitigating factor while "she is pretending to be a battered woman" is not).

We chose to conduct a manual analysis because prior work by Wen-Yi et al. (2024) found that automating such analyses using language models was ineffective, and often lost case-specific nuances captured by expert review. The coding team consisted of the first two authors, who coded 85 passages, representing 44 men's transcripts and 40 women's transcripts.

4 Results

Many of the themes we identified consistently appeared more frequently in women's transcripts compared to men's. Most prominent are *Gender Roles*, *Sexualization*, *Appearance*, and *LGBTQ+ Identity*; other themes include *Characterization*, *Emotions*, *Parenting*, *Intimate Relationships*, *Poverty & Spending*, and *Religion*. A small number of themes generally appeared more frequently in men's transcripts compared to women's. These include: *Drugs & Addiction*, *Racial Tropes*, *Immigration*, and *Violence*.

Two caveats shape how the results should be read. First, the presence of a key term does not, in and of itself, establish stereotyping; additional analysis (such as the qualitative analysis we employ) is needed to determine whether a key term is used in a stereotyping way. Second, because our themes and key terms were selected based on a review of the literature on gender bias against women defendants specifically, our quantitative results to some extent reflect the asymmetry in our analytic instrument. However, the comparison provides evidence that women and men defendants are spoken about differently, on average. We suggest that stereotyping may arise as a result of these disparate descriptions. Similarly, the fact that key terms related to *Drugs & Addiction*, *Racial Tropes*, *Immigration*, and *Violence* occur more frequently in men's transcripts does not imply that women are not subject to the biases related to these themes, but rather that the terms used in women's transcripts may be substantively different from, and perhaps less well-documented than, those used in men's transcripts.

Our results are summarized at the theme level in Table 2 and OSF Appendix Figure A1. Individual key terms that are relatively more common in women's and men's transcripts are shown in OSF Appendix Tables A4–A11. In the remainder of this section, we describe results in detail at the matched group level. We focus primarily on our large-scale quantitative analysis and supplement findings with our small-scale qualitative analysis. We anonymize all quotes and details sourced from our corpus in this section (unless the quotes or details have already been made public) to preserve the privacy of defendants and of victims.

Theme	Crime Match	Parent Defendant Match	Child Victim Match	Intimate Partner Victim Match	Defendant Race Match (White)	Defendant Race Match (Black)	Defendant Race Match (Latine)	Defendant Race and Crime Match
<i>Gender Roles</i>	W	W	W	W	W	W	W	W
<i>Sexualization</i>	W	W	-	W	W	W	W	W
<i>Appearance</i>	W	W	-	-	W	-	-	W
<i>Characterization</i>	W	W	-	W	W	W	M	W
<i>Emotions</i>	W	W	M	W	W	W	W	W
<i>Parenting</i>	W	W	M	W	W	W	W	W
<i>Intimate Relationships</i>	W	W	M	W	W	W	W	W
<i>LGBTQ+ Identity</i>	W	W	-	W	W	-	W	W
<i>Mental Health & Disability</i>	W	M	M	W	-	-	W	W
<i>Drugs & Addiction</i>	M	-	M	-	-	M	M	M
<i>Poverty & Spending</i>	W	W	W	-	W	M	W	W
<i>Racial Tropes</i>	M	M	-	-	M	M	M	M
<i>Immigration</i>	-	M	M	M	M	-	M	M
<i>Religion</i>	W	-	M	W	-	-	-	-
<i>Metaphor</i>	M	-	W	W	M	M	W	M
<i>Violence</i>	-	-	M	M	M	-	-	M

Table 2: Summary of theme-level results. Themes whose associated key terms appear statistically more commonly in women’s transcripts ($p < 0.05$) are noted with a ‘W’ and colored in red; themes whose associated key terms appear statistically more commonly in men’s transcripts are noted with an ‘M’ and colored in blue. Themes with practically significant differences (absolute difference > 0.05) are bolded and have darker-colored backgrounds.

Crime Match. Among matched transcripts of defendants who had the same aggravating factors added to their murder charge, we observe practically significant differences in the presence of terms related to *Sexualization*, *Parenting*, *Intimate Relationships*, and *LGBTQ+ Identity* (more common for women); as well as *Racial Tropes* (more common for men). This is in line with prior work that has found that women’s sexual history, sexuality, and parenting are emphasized by prosecutors to convict women even when such details may be irrelevant to their alleged crime (Babcock and Greenfield 2023; Babcock, Greenfield, and Adamson 2024; Greenfield and Babcock 2026; Potts and Weare 2018). Our findings provide empirical evidence for these findings at a large scale.

A qualitative analysis of selected passages also reveals substantive differences in how key terms associated with a

single theme are used across gender. For example, *Mental Health & Disability* in women is discussed using a wide variety of terms (e.g., “psychosis”, “bipolar”, “medications”); the term “syndrome” (as in “battered woman syndrome”) is frequently used. In contrast, references to men’s mental health are more likely to reference key terms like “brain,” “IQ,” and “smart.” Across defendant genders, we find that defense attorneys raise mental health as influencing defendants’ behavior, in some cases to explicitly argue against a death sentence. Prosecutors seek to undermine such arguments, often using expert testimony to argue in favor of mental competency of the defendant. Men’s cases tended to argue within the framework of IQ levels, while women’s cases tended to refer to specific diagnoses. Language related to mental health is particularly important because researchers estimate that at least 40% of people on

death row have had a serious mental illness prior to their criminal charge (Baumgartner et al. 2017; Hunt 2023). In the U.S., it is unconstitutional to impose the death penalty upon individuals with an intellectual disability (DPI 2025b). However, in part because of the absence of a uniform definition, prosecutors may selectively minimize or weaponize mental illness or disability during capital trials (Farr 2022; Hunt 2023). Much of the prior literature on this topic focuses on cases involving men (Baumgartner et al. 2017; Hunt 2023). We find that terms related to *Mental Health & Disability* are more varied in the women's transcripts in our corpus, suggesting that more research is needed into the particular ways in which the mental health of women defendants specifically is discussed in capital trials.

Parent Defendant Match. Among the transcripts of defendants who are parents, we find that terms related to *LGBTQ+ Identity* and *Poverty & Spending* occur more frequently for women. *Parenting* language is also statistically significantly more common in women's transcripts—despite the fact that all defendants in this sample are parents. This result appears directionally consistent with prior findings that women are more severely punished and negatively characterized in relation to their status as parents (Potts and Weare 2018).

We find qualitative evidence that, for men, *Parenting* language often refers to the defendant's background, including their childhood or their relationship to family members, rather than their own status as parents. Conversely, *Parenting* language is used to describe women's characteristics as mothers. Interestingly, we also note cases where *Parenting* language is not used directly in reference to women defendants, but may still be used against them: for example, during the trial of a woman defendant for the murder of her niece, prosecutors referred to the defendant's sister (the mother of the deceased child) as a "horrible mother" who "shouldn't have children" in an effort to discredit her testimony on behalf of the defense.

Child Victim Match. The transcripts of women and men accused of murdering their own child (filicide) are more similar across the themes we analyzed than any other grouping we consider. Unlike other capital murder crimes, filicide is estimated to be perpetrated roughly equally by women and men (Dawson 2015). In contrast to all other groupings, *Parenting* language appears statistically more frequently in the transcripts of men than of women accused of killing their child. We interpret this as providing further support for the hypothesis that women who are parents face stigmatization based on their perceived failures as parents even when their parental status is not relevant to their crime—because when parental status *is* relevant to the crime, we see that *Parenting*-associated key terms are no longer used more frequently for women in our corpus.

Intimate Partner Victim Match. Women and men accused of murdering a current or former romantic partner had some of the largest differences in language across transcripts. Terms related to *Gender Roles*, *Emotions*, *Parenting*, *Intimate Relationships*, *LGBTQ+ Identity*, *Mental Health*

& *Disability*, and *Religion* are more frequent in women's transcripts; terms related to *Violence* are more frequent in men's. Again, we see *Parenting* language used in women's transcripts even though it is not necessarily relevant to the alleged crime; similarly, we see language related to *LGBTQ+ Identity* used more frequently for women, in line with prior work (Farr 2000). An analysis of our collected metadata shows that women defendants are more likely to have a current intimate relationship with the victim (9/9 in our sample), while men defendants are more likely to have a prior intimate relationship with the victim (3/6 in our sample), a difference that may partially account for the very different ways that language is used in the transcripts of women and men in this sample.

Qualitative analysis shows how language related to *Religion* is used specifically in the transcripts of women accused of murdering an intimate partner, although gender-based differences in references to *Religion* are rare across the other matched samples. For example, during the trial of a woman defendant accused of murdering her husband, prosecutors emphasized that the defendant "betrayed her Christian upbringing."

Race Match: White Defendants. Within the transcripts of white defendants, we find that language related to *Sexualization* is used in women's transcripts more often than in men's. Interestingly, we find that *Racial Tropes* are more frequently used in cases of white men compared to white women. A term-level examination shows that this is due to the use of words like "gang," which have a racialized connotation but can also be used against white men defendants.

Race Match: Black Defendants. Within the transcripts of Black defendants, we find that language related to *Emotions*, *Parenting*, and *Intimate Relationships* is used in women's transcripts more often than in men's, underscoring how Black women face heightened vulnerability to gendered family and parenting stereotypes in court (Franklin 2013; Rosenthal and Lobel 2016). The most disproportionately common terms in women's transcripts include neutral terms like "baby" and "child" as well as abuse-related terms like "belt" and "bruise." In contrast, the most disproportionately common terms in men's transcripts are related to *Drugs & Addiction* ("cocaine," "sell*", "pipe") and *Racial Tropes* ("gang").

Race Match: Latine Defendants. Within the transcripts of Latine defendants, we find the largest disparities in terms related to *Gender Roles*, *Emotions*, *Intimate Relationships*, and *Metaphor*, all of which occur more frequently in women's transcripts. Language related to *Drugs & Addiction* and *Racial Tropes*, including explicit references to nationality (e.g., "Mexican"), occurs more frequently in the men's transcripts.

Race and Crime Match. Finally, within the transcripts of defendants matched both on their race and on the similarity of their crimes, we again find that language invoking *Sexualization* and *LGBTQ+ Identity* is more frequently used in women's transcripts, while language invoking *Racial Tropes* is, once again, more frequently used in men's. Overall,

we demonstrate the consistency of these trends even when accounting for multiple potentially confounding variables.

Qualitatively, we find that language in this sample invokes *Appearance* and *Characterization* in ways that could discredit women defendants (key terms associated with both themes are used statistically more frequently in the transcripts of women in this sample). For example, a woman defendant accused of murdering her friend was described as “very fake” because she did not like to go out without “some kind of hairstyle.” Prosecutors in the trial of one woman defendant accused of murdering her children emphasized her “vanity;” prosecutors in the trial of another critiqued the defendant for “laughing and giggling” at a posthumous birthday party for one of her sons in order to call her character into question.

5 Discussion

In this section, we discuss the implications of our findings for the U.S. criminal punishment system as well as for the academic field of legal NLP. We also outline limitations and suggest avenues for future work.

Implications for the U.S. Criminal Punishment System.

Most importantly, our results have important implications for the defendants in our corpus, and for other individuals currently on death row. Although our analysis was simple, it revealed language that is indicative of potentially prejudicial gender stereotyping across a large number of women’s capital trial transcripts. As the Supreme Court ruled in *Andrew v. White* (2025), the prosecution’s utilization of that language could provide a basis for arguing that a trial was “fundamentally unfair,” which might in turn provide an avenue for reconsideration of a conviction or death sentence. Overall, our findings support prior abolitionist scholarship in underscoring the discriminatory application and fundamentally unfair nature of the death penalty in the U.S. (Foran, Kaba, and Naples-Mitchell 2021; Mogul 2005).

Implications for Legal NLP. Several legal NLP applications use transcripts as training data, or as input for prediction tasks (Siino et al. 2025). One such task is legal document summarization, which entails providing concise summaries of legal texts (Feijo and Moreira 2023; Kanapala, Pal, and Pamula 2019). Another is legal judgment prediction (LJP), which uses legal documents to predict the judgment in a case (Kumar 2024; Nigam et al. 2024; Siino et al. 2025; Valvoda and Cotterell 2024). As Medvedeva and McBride (2023) highlight, LJP accuracy can vary widely based on how the ‘facts’ of a case are articulated and which ‘facts’ are reported. Our findings raise important considerations for these tasks. In the context of legal document summarization, terms that are repeated frequently are more likely to be seen as important and captured by and emphasized in the summary. Gender-associated key terms, which are often repeated throughout trials, could therefore be captured in document summaries. Similarly, to the extent that LJP relies on subjective summaries of the ‘facts’ of a case, which are often constructed by judges through review of court transcripts, repeated use of gender-associated key terms

may affect the outcomes of LJP models. This is important in practice because LJP models are often used by courts to prioritize cases or by law firms to decide which cases to accept. Biases in transcripts therefore may have downstream impacts on how cases are prioritized and the representation that defendants have access to.

Finally, the presence of gender-associated key terms in capital trials calls the use of AI in the criminal punishment system into question. As foundation models are increasingly adopted for legal work, legal texts—including trial transcripts—are being used to train and fine-tune models for legal tasks (Ariai, Mackenzie, and Demartini 2024; Nikita, Rana, and Mehta 2023). The gender-associated key terms we identify in this study can influence the outcomes of these tasks, potentially propagating gender bias into future legal proceedings.

Limitations. Our study is subject to several limitations. While our key terms were carefully selected through an iterative process that leveraged domain expertise, NLP analysis of transcripts, and expert data annotations, the set of terms cannot possibly represent all ways of expressing a gender stereotype. The list we develop is ultimately subjective and non-exhaustive, and is focused primarily on women-associated terms. Our crime groupings are also subjective—there are other reasonable ways to group aggravating factors that could differ from ours. This could result in slightly different matched groups. Another data-related limitation is that a small number (six) of women’s trials involved women who were jointly tried with male co-defendants, and therefore include language in reference to men defendants as well as women defendants. Finally, while we offer an analysis of terms related to LGBTQ+ identity, we were unable to collect metadata on the sexual orientation of individuals in our corpus, limiting our analysis of how this language may play a role specifically in the trials of LGBTQ+ individuals.

In terms of methodological limitations, we employ a simple term frequency-based NLP method that is unable to capture the nuances of coded and more subtle language. Another methodological limitation is our approach of bootstrapping 500 word ‘documents’ rather than full transcripts, which we chose to do to mitigate the volatility in document-level analysis for small datasets. While this improves the statistical power of our bootstrap estimates, it introduces a violation of the traditional assumption that documents are independent and identically distributed. As a result, the inflated sample size artificially narrows our confidence intervals and deflates *p*-values, potentially overstating the observed statistical significance. To mitigate this issue, we focus our results largely on practical significance (i.e., large absolute differences in point estimates) rather than relying solely on statistical significance.

Future Work. Our study identifies the presence of gendered discourse based on the prevalence of gender-associated key terms. Future work can build on this template in several ways. First, future work can incorporate more complex NLP techniques to computationally identify the speaker, subject, and sentiment of a sentence in a transcript instead of relying solely on qualitative analysis,

as we did in this work. This could be used to further narrow down portions of transcripts for legal teams to focus on for detailed analysis. Second, future work could incorporate techniques beyond term match to identify stereotyping that manifests through subtle or coded language. Finally, future work can explore the extent to which gendered discourse has been used in the trials of men defendants on death row by applying an analogous approach to ours.

6 Conclusion

In this work, we present what is to our knowledge the largest-scale examination of gender-biased terms in U.S. capital trials to date. We use simple and interpretable NLP techniques to analyze 137 capital trial transcripts, including the trials of nearly every woman currently on death row. We find that, even when conditioning on defendant demographics and crime characteristics, terms associated with *Gender Roles*, *Sexualization*, *Appearance*, and *LGBTQ+ Identity* are used more frequently in the trials of women; terms associated with *Characterization*, *Emotions*, *Parenting*, *Intimate Relationships*, *Poverty & Spending*, and *Religion* are usually used more frequently in the trials of women. When there are exceptions—for example, *Parenting* language is more commonly used against men accused of killing their child than women accused of doing so—they provide empirical evidence to show that negative tropes about defendants being “bad mothers” are used against women even when *Parenting* is not relevant to their case. Similarly, we find that even among defendants convicted of similar crimes, language related to *Sexualization* remains more widely deployed in women’s trials, suggesting this language is not merely a reflection of case facts. Through a small-scale qualitative analysis enabled by our computational findings, we find evidence supporting our hypothesis that the presence of gender-associated key terms can indicate the presence of stereotyping. We also identify differences in the use of terms associated with the same theme across men’s and women’s transcripts. While these results should not be taken to imply that gender stereotyping more commonly affects women than men, they do provide evidence that women and men defendants are spoken about differently, on average, in capital trials. Our results affirm prior work (e.g., Potts and Weare 2018; Babcock and Greenfield 2023; Greenfield and Babcock 2026) that highlights stereotyping against women in U.S. capital trials. Based on our findings, we call for additional legal scholarship focused specifically on women capital trial defendants.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Number 2416996. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1972. *Furman v. Georgia*. 408 U.S. 238 (1972).
1976. *Gregg v. Georgia*. 428 U.S. 153 (1976).
1987. *McCleskey v. Kemp*. 481 U.S. 279 (1987).
2025. *Andrew v. White*. 604 U.S. ____ (2025).
- Aley, M.; and Thomas, B. 2021. An examination of differences in product types and gender stereotypes depicted in advertisements targeting masculine, feminine, and LGBTQ audiences. *Communication Research Reports*, 38(2): 132–141.
- Ammons, L. L. 1995. Mules, madonnas, babies, bathwater, racial imagery and stereotypes: The African-American woman and the battered woman syndrome. *Wis. L. Rev.*, 1003.
- Andina-Diaz, E.; Ventura-Miranda, M. I.; Quiroga-Sanchez, E.; Ortega-Galan, A. M.; Fernandez-Medina, I. M.; and Ruiz-Fernandez, M. D. 2023. Nursing Students’ Perception about Gender Inequalities Presented on Social Networks: A Qualitative Study. *International Journal of Environmental Research and Public Health*, 20(3).
- Ariai, F.; Mackenzie, J.; and Demartini, G. 2024. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. *arXiv preprint arXiv:2410.21306*.
- Atwell, M. W. 2007. *Wretched sisters: Examining gender and capital punishment*, volume 19. Peter Lang.
- Avery, J. J. 2019. Addiction stigma in the US legal system. In *The stigma of addiction: An essential guide*, 153–166. Springer.
- Babcock, S. 2024. Gendered Capital Punishment. *William & Mary Journal of Race, Gender, and Social Justice*, 31.
- Babcock, S.; and Greenfield, N. 2023. Gender, Violence, and the Death Penalty. *California Western International Law Journal*, 53(2): 327–403.
- Babcock, S.; Greenfield, N.; and Adamson, K. 2024. Women on Death Row in the United States. *Cardozo Law Rev.*, 46: 1.
- Baldus, D. C.; Pulaski, C.; and Woodworth, G. 1983. Comparative review of death sentences: An empirical study of the Georgia experience. *The Journal of Criminal Law and Criminology* (1973-), 74(3): 661–753.
- Basow, S. A. 1992. *Gender: Stereotypes and roles*. Thomson Brooks/Cole Publishing Co.
- Baumgartner, F.; Davidson, M.; Johnson, K.; Krishnamurthy, A.; and Wilson, C. 2017. *Deadly justice: A statistical portrait of the death penalty*. Oxford University Press.
- Benatti, R.; Severi, F.; Avila, S.; and Colombini, E. L. 2024. Gender Bias Detection in Court Decisions: A Brazilian Case Study. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’24, 746–763. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.
- Bjerregaard, B.; Weber, C.; Smith, M. D.; Cochran, J. K.; and Fogel, S. J. 2026. An Empirical Analysis of the Heinous, Atrocious, and Cruel (HAC) Aggravating Circumstance: A Focal Concerns Approach. *American Journal of Criminal Justice*.

- Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; and Kalai, A. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, NIPS. ArXiv:1607.06520 [cs, stat].
- Bosco, C.; Patti, V.; Frenda, S.; Cignarella, A. T.; Paciello, M.; and D’Errico, F. 2023. Detecting racial stereotypes: An Italian social media corpus where psychology meets NLP. *Information Processing & Management*, 60(1): 103118.
- Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2): 77–101.
- Bright, S. B. 1994. Counsel for the Poor: The Death Sentence Not for the Worst Crime but for the Worst Lawyer. *Yale Law Journal*, 103: 1835–1883.
- Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.
- Capelle, M.; Frasinicar, F.; Moerland, M.; and Hogenboom, F. 2012. Semantics-based news recommendation. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, WIMS ’12. New York, NY, USA: Association for Computing Machinery. ISBN 9781450309158.
- Carll, E. K. 2003. News portrayal of violence and women: Implications for public policy. *American Behavioral Scientist*, 46(12): 1601–1610.
- Cignarella, A. T.; Giachanou, A.; and Lefever, E. 2025. A Survey on Stereotype Detection in Natural Language Processing. *ACM Comput. Surv.*, 58(5).
- Cignarella, A. T.; Sanguinetti, M.; Frenda, S.; Marra, A.; Bosco, C.; and Basile, V. 2024. QUEEREOTYPES: A multi-source Italian corpus of stereotypes towards LGBTQIA+ community members. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 13429–13441.
- Cornell Center on the Death Penalty Worldwide. 2018. Judged for more than her crime: A global overview of women facing the death penalty.
- Cuddy, A. J.; Fiske, S. T.; and Glick, P. 2008. Warmth and Competence as Universal Dimensions of Social Perception: The Stereotype Content Model and the BIAS Map. In *Advances in Experimental Social Psychology*, volume 40, 61–149. Elsevier. ISBN 978-0-12-015240-7.
- Dawson, M. 2015. Canadian trends in filicide by gender of the accused, 1961–2011. *Child abuse & neglect*, 47: 162–174.
- Death Penalty Information Center. 2025a. Death Row: Women.
- Death Penalty Information Center. 2025b. Executions Overview.
- Deshpande, A.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; and Narasimhan, K. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.
- Easteal, P.; Bartels, L.; Nelson, N.; and Holland, K. 2015. How are women who kill portrayed in newspaper media? Connections with social values and the legal system. *Women’s Studies International Forum*, 51: 31–41.
- Farr, K. 2022. Intellectual disability and mental illness among women sentenced to death in the US: Constitutional and evidentiary dilemmas. *Punishment & Society*, 24(4): 571–591.
- Farr, K. A. 2000. Defeminizing and dehumanizing female murderers: Depictions of lesbians on death row. *Women & Criminal Justice*, 11(1): 49–66.
- Fast, E.; Vachovsky, T.; and Bernstein, M. 2016. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, 112–120.
- Feijo, D. d. V.; and Moreira, V. P. 2023. Improving abstractive summarization of legal rulings through textual entailment. *Artificial intelligence and law*, 31(1): 91–113.
- Fiske, S.; Cuddy, A.; Glick, P.; and Xu, J. 2002. A model of stereotype content as often mixed: Separate dimensions of competence and warmth respectively follow from status and competition. *Journal of Personality and Social Psychology*, 82(6): 878–902.
- Fiske, S. T. 1998. *The handbook of social psychology*, chapter Stereotyping, prejudice, and discrimination. McGraw-Hill, 4th edition.
- Flanders, C. 2019. What Makes the Death Penalty Arbitrary? And Does it Matter if it is? *Wisconsin Law Review*, 2019(1).
- Foran, R.; Kaba, M.; and Naples-Mitchell, K. 2021. Abolitionist Principles for Prosecutor Organizing: Origins and Next Steps. *Stanford Journal of Civil Rights and Civil Liberties*, 16(3): 496–536.
- Franklin, T. W. 2013. Sentencing Native Americans in US federal courts: An examination of disparity. *Justice Quarterly*, 30(2): 310–339.
- Fraser, K. C.; Kiritchenko, S.; and Nejadgholi, I. 2022. Computational modeling of stereotype content in text. *Frontiers in artificial intelligence*, 5: 826207.
- Fraser, K. C.; Kiritchenko, S.; and Nejadgholi, I. 2024. How Does Stereotype Content Differ across Data Sources? In *Proceedings of the 13th joint conference on lexical and computational semantics (*SEM 2024)*, 18–34.
- Fraser, K. C.; Nejadgholi, I.; and Kiritchenko, S. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. *arXiv preprint arXiv:2106.02596*.
- Greenfield, N.; and Babcock, S. L. 2026. Sex on Trial. *Columbia Journal of Gender and Law*, 46(3): 219–294.
- Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
- Hunt, A. 2023. Declining Competency: Protecting Defendants with Worsening Mental Illness on Death Row from the Death Penalty. *BCL Rev.*, 64: 1723.

- Jenkins, P.; and Davidson, B. 1990. Battered women in the criminal justice system: An analysis of gender stereotypes. *Behavioral Sciences & the Law*, 8(2): 161–170.
- Jones, S. 2022. Ending Extreme Sentencing is a Women’s Rights Issue. *Georgetown Journal of Gender and the Law*, 23.
- Kanapala, A.; Pal, S.; and Pamula, R. 2019. Text summarization from legal documents: a survey. *Artificial Intelligence Review*, 51(3): 371–402.
- Kumar, M. 2024. Predictive modelling in legal decision-making: Leveraging machine learning for forecasting legal outcomes. *J. Electrical Systems*, 20(3): 2060–2071.
- LeBeau, K.; Carr, C.; and Hart, M. 2020. Examination of gender stereotypes and norms in health-related content posted to Snapchat discover channels: qualitative content analysis. *Journal of Medical Internet Research*, 22(3): e15330.
- Levin, B. 2023. After The Criminal Justice System. *Washington Law Review*, 98(3): 899–946.
- Locatelli, D.; Damo, G.; and Nozza, D. 2023. A cross-lingual study of homotransphobia on Twitter. In *Proceedings of the First Workshop on Cross-cultural Considerations in NLP (C3NLP)*, 16–24.
- Lucy, L.; Demszky, D.; Bromley, P.; and Jurafsky, D. 2020. Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in Texas US history textbooks. *AERA Open*, 6(3): 2332858420940312.
- Medvedeva, M.; and McBride, P. 2023. Legal judgment prediction: If you are going to do it, do it right. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, 73–84.
- Mendelsohn, J.; Tsvetkov, Y.; and Jurafsky, D. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in artificial intelligence*, 3: 55.
- Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781.
- Mogul, J. L. 2005. The dykier, the butcher, the better: The state’s use of homophobia and sexism to execute women in the United States. *NY City L. Rev.*, 8: 473.
- Nigam, S. K.; Deroy, A.; Maity, S.; and Bhattacharya, A. 2024. Rethinking Legal Judgement Prediction in a Realistic Scenario in the Era of Large Language Models. In Aletras, N.; Chalkidis, I.; Barrett, L.; Goantă, C.; Preotiuc-Pietro, D.; and Spanakis, G., eds., *Proceedings of the Natural Legal Language Processing Workshop 2024*, 61–80. Miami, FL, USA: Association for Computational Linguistics.
- Nikita; Rana, D. P.; and Mehta, R. G. 2023. Research Challenges for Legal Document Summarization. In *2023 IEEE World Conference on Applied Intelligence and Computing (AIC)*, 307–312.
- Pierce, G. L.; Radelet, M. L.; and Sharp, S. 2017. Race and Death Sentencing for Oklahoma Homicides Committed Between 1990 and 2012. *The Journal of Criminal Law & Criminology*, 107: 733–756.
- Pinto, A. G.; Cardoso, H. L.; Duarte, I. M.; Warrot, C. V.; and Sousa-Silva, R. 2020. Biased language detection in court decisions. In *International Conference on Intelligent Data Engineering and Automated Learning*, 402–410. Springer.
- Potts, A.; and Weare, S. 2018. Mother, monster, Mrs, I: A critical evaluation of gendered naming strategies in English sentencing remarks of women who kill. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*, 31(1): 21–52.
- Poznanski, J.; Soldaini, L.; and Lo, K. 2025. olmOCR 2: Unit Test Rewards for Document OCR. arXiv:2510.19817.
- Rosenthal, L.; and Lobel, M. 2016. Stereotypes of Black American women related to sexuality and motherhood. *Psychology of women quarterly*, 40(3): 414–427.
- Sheng, E.; Chang, K.-W.; Natarajan, P.; and Peng, N. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, EMNLP, 3405–3410. Hong Kong, China: Association for Computational Linguistics.
- Siino, M.; Falco, M.; Croce, D.; and Rosso, P. 2025. Exploring llms applications in law: A literature review on current legal nlp approaches. *IEEE Access*, 13: 18253–18276.
- Slakoff, D. C. 2020. The representation of women and girls of color in United States crime news. *Sociology Compass*, 14(1): e12741.
- Sutton, J.; Mills, J.; Merrigan, J.; and Swain, K. 2021. Death By Dehumanization: Prosecutorial Narratives of Death-Sentenced Women and LGBTQ Prisoners. *St. John’s Law Review*, 95(4): 1053–1071.
- The Legal Defense Fund. 2025. DEATH ROW U.S.A.: Fall 2025.
- Valvoda, J.; and Cotterell, R. 2024. Towards Explainability in Legal Outcome Prediction Models. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 7269–7289. Mexico City, Mexico: Association for Computational Linguistics.
- Wen-Yi, A. W.; Adamson, K.; Greenfield, N.; Goldberg, R.; Babcock, S.; Mimno, D.; and Koenecke, A. 2024. Automate or assist? The role of computational models in identifying gendered discourse in US capital trial transcripts. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1556–1566.