PROJECT RECON:
A COMPUTATIONAL FRAMEWORK FOR AND ANALYSIS OF THE CALIFORNIA PAROLE HEARING SYSTEM

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF MANAGEMENT SCIENCE AND ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Jenny Hong
March 2023

This dissertation is online at: https://purl.stanford.edu/xn213ms8118

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Johan Ugander, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Christopher Manning, Co-Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Itai Ashlagi

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Sharad Goel

Approved for the Stanford University Committee on Graduate Studies.

Stacey F. Bent, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format.

Preface

Parole decisions can tip a sentence toward fifteen years or fifty. Despite the great power that parole boards hold, their decision processes are poorly documented and largely hidden from public scrutiny. Parole hearings produce almost no structured data, only an unstructured transcript of hearing dialogue several hundred pages in length. In the following dissertation, we use natural language processing to analyze the transcripts of 35,105 parole hearings held between 2007 and 2019 for candidates serving life sentences in California, totalling approximately five million pages. Through regression analyses of data extracted from the transcripts, after controlling for relevant case factors, we find that several factors outside of the candidate's control explain hearing outcomes. We find that commissioners vary widely in their punitiveness in previously unobserved ways; the assignment to a particular commissioner significantly influences the hearing outcome. Racial disparities limit the quality of legal representation that parole candidates receive as well as their voice in the hearing dialogue, and both significantly predict the parole outcome after again controlling for case factors. Previous analyses of parole systems have been limited by the unavailability of structured data or the task of hand-annotating hearing transcripts. Our results thus provide the most comprehensive picture of a parole system studied to date. While our results carry direct implications for legislative parole reform, our methodology—using machine learning to analyze legal hearings—can be extended to many other procedures in criminal and administrative law with limited structured data.

Acknowledgments

I am deeply grateful for my co-advisers Johan Ugander and Chris Manning, not only for their insights and intellectual contributions, but also for their patience and encouragement. Thank you for taking on a project that had been born as an academic foster child, and for joining me in building it a home. I remember clearly the challenges to the project that you foresaw in my first year, and I can't overstate what an honor it has been for you to face those challenges with me. I would not be the researcher I am today without both of you.

The inspiration for Project Recon originated from the conversations, collaborations, and mentorship of Nick McKeown and Kristen Bell, and their professional and extra-professional dedication to criminal justice. Their work, and that of the Stanford Criminal Justice Center, the Berkeley Law Death Penalty Clinic, UnCommon Law, No More Tears, and so many more, remind me that this dissertation is not an act of my passing on a torch; I have never been the first, last, or only one on this path. Thank you to Larry Rosser, Miguel Quezada, Lonnie Morris, Keith Wattley, Isaac Dalke, and everyone else whose parallel journeys have shaped this dissertation.

Thank you to the Electronic Frontier Foundation for representing the Project Recon team in our litigation for the public records used in this dissertation. In spite of the impasse that litigation posed to the dissertation, it was also an opportunity to distill our values on the record and to communicate the importance of transparency and fairness.

I thank my committee and the broader academic community I have found at Stanford. Conversations with Sharad Goel and Itai Ashlagi and their students for their intellectual contributions to this dissertation. I am incredibly lucky to have counted myself as part of the NLP group and the ULab, both of which have greatly enriched my PhD experience with a breadth of intellectual ideas and a welcoming social group. The meetings, seminars, and social events with both of these groups were often the highlight of my week.

I am grateful to the research assistants at the University of Oregon and at Stanford for their varied contributions to Project Recon. This dissertation does not capture the breadth of ideas that we were able to collectively explore thanks to the help of many creative individuals. I would like to thank Derek Chong and Graham Todd in particular for their initiative in taking ownership of specific aspects of Project Recon.

By far, the one person I have had the privilege of working the most closely with over the course of my PhD is Catalin Voss. For the first several years of my PhD, Catalin and I were practically joined the hip, taking equal share and ownership of Project Recon, and I could not have asked for more of a research partner. His creativity, energy, and optimism were my light through many weekends and late nights stuck on technical or bureaucratic problems. He was the one who kept me coming back to the office day after day.

As formative of an experience as the PhD has been, at times it feels like I ended up in this program by chance. I applied to the PhD based on the suggestion of a single mentor, Stephen Boyd, who I am grateful to have worked with and who suggested that such an endeavor might be something I could reconcile with my interests in social justice. And it was Raja GuhaThakurta who first inspired in me the awe for scientific discovery. I thank both Stephen and Raja for inspiring my love of teaching and mentorship. I thank Patrick Valencia and Stuart Scott for giving me the necessary hope for the future to choose a PhD program.

Finally, thank you to the village that it took to raise me. The earliest supporters of my education were those who themselves did not have the opportunity to learn how to read. To all those who have put a roof over my head, food on the table, and love in my heart, thank you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The United States has held the record for the highest incarceration rate in the world for several decades, with between 600 and 800 individuals per hundred thousand, outpacing countries such as Turkmenistan and Rwanda, and exceeding the incarceration rate of neighboring countries, Mexico and Canada, by roughly six times [Walmsley, 2003, Fair and Walmsley, 2021, Widra and Herring, 2021]. Given the size of its population, the United States has also held the record for the total number of incarcerated individuals, which at the time of writing is just over two million individuals. It is no surprise that the criminal justice system and its constituent components have been an active area for litigation, activism, and scholarship.

In recent years, criminal justice reform has reduced the rate of new incarceration across many states. However, the recent modest decline in America's prison population has not compensated for four decades of policies that actively promote incarceration [Ghandnoosh, 2020]. One of the counteracting mechanisms in the American criminal justice system is parole, the conditional release from prison, which provides a direct mechanism for states to meaningfully relieve the pressure on their overcrowded prisons [Reitz and Rhine, 2020].

Parole can be the deciding factor between whether a sentence lasts fifteen years or fifty. In many cases, the timing of when someone receives parole can determine a sentence length even more than the initial sentencing hearing.

Despite the unique role that parole plays in the criminal justice system, there has been no large scale analysis of parole to date. The scarcity of quantitative analyses of parole does not reflect a lack of quantitative analyses of the criminal justice system. Such quantitative and often large-scale studies comprise a growing body of legal scholarship. In a sense, empirical methods have simply "crept into" the publications of legal scholars [Ulmer, 1963]. In more recent years, however, the fields of Empirical Legal Studies, as well as many "law and" fields, such as law and economics or law and sociology, have actively welcomed contributions from scholars across economics, psychology, health care, policy, political science, criminology, finance, and sociology [Eisenberg, 2011].

Within the criminal justice system, a number of scholars have undertaken quantitative studies on various components of the system, such as policing [Gelman et al., 2007, Pierson et al., 2020], bail [Arnold et al., 2018], and sentencing [Klein et al., 1990, Abrams et al., 2012], each building on decades of existing, often smaller-scale empirical studies of the same components. In other words, most of the empirical attention has focused on the many decision steps involved in how individuals *reach* prison. Comparatively little has focused on how someone *leaves* the prison system.

In part, parole has been hidden from public view because, broadly, parole falls in the domain of prisons, which operate with little transparency. There are 52 parole boards in the United States,[1] most of which are operated by departments of corrections. The parole boards are the custodians of the data that researchers require to understand parole, and the boards have made that data difficult, or in some cases impossible, to obtain. Existing studies of parole have been limited to the study of a handful of tabular features [Weisberg et al., 2011, Friedman and Robinson, 2014, Young, 2016], such as demographic data, or, where transcripts of dialogue are available, meticulous hand-annotation of a small sample of 107–754 transcripts [Bradley and Engen, 2016, Bell, 2019, Greene and Dalke, 2020].

One lens through which to view the present dissertation is as an empirical study of a legal domain that contributes the *first* large-scale study of a parole system through an analysis of a complete historical record of all 35,105 California parole transcripts from 2007–2019. This dissertation answers a range of questions about the California parole hearing system, ranging from basic tabulations of parole grant rates to linguistic analyses of how parole attorneys represent their candidates. We call this the *domain application lens*.

However, viewing the dissertation only through the lens of a domain contribution to parole fails to address the question of timeliness. What enables the scale of this research, which draws its insight from a dataset approximately *one hundred* times larger than prior studies of California parole hearings? What technological tools have been developed and applied to this research that were not available to prior studies?

The second lens through which to view this research is the *machine learning lens*, as a contribution to the ongoing application of new Natural Language Processing (NLP) capabilities to various domain applications. The primary NLP challenge addressed in this dissertation is the process of extracting structured data from parole hearing transcripts for use in downstream analysis. Much of the research effort within NLP is directed toward teaching machines to perform a particular task. The challenges in this dissertation closely relate to, but are not fully solved by, several such tasks, such as Multiple-Choice Reading Comprehension and Open-Domain Question Answering.[2]

Despite the many different ways our question could be modeled, two challenges are shared

[1] Each of the fifty states has a parole board, and the U.S. Parole Commission and the Naval Clemency and Parole Board serve as two additional parole boards.

[2] Section 6.2 defines and further explores the appropriateness of various tasks for modeling challenges in parole hearings.

across different choices of tasks. The first challenge is that of document length. Transformer models [Vaswani et al., 2017] have achieved state-of-the-art results across a range of tasks both within and beyond NLP. However, transformer models do not scale up well to read longer inputs. They require an amount of computation that is quadratic relative to the length of the input. That is, doubling the length of a document requires four times as much computation, tripling the length of a document requires nine times as much computation, and so on. Most transformer models can reasonably process documents of approximately five hundred to one thousand words.[3] The average number of words in a parole hearing in our study is twenty thousand words, which is out of the range of most transformer models. Extending the input length of transformer models has been an area of active research [Child et al., 2019, Roy et al., 2021, Kitaev et al., 2020, Zaheer et al., 2020], even as for present NLP benchmarks, longer contexts are often not well utilized.

The second challenge is the difficulty of generating training data for the parole hearing task. Because of the specialized nature of the parole domain, we rely on Subject Matter Experts (SMEs) to annotate parole hearings to use as training examples for a language model. Even after SMEs are identified and trained, they still require a large amount of time to annotate one hearing. This motivates research into best practices for training models with sparse training data, and also into other methods for incorporating SME knowledge. Toward the latter goal, a framework known as data programming [Ratner et al., 2016] suggests an alternative to annotating parole hearings one at a time.

Long input documents and the scarcity of training data are two challenges shared among many applications within natural legal language processing (NLLP), a growing interdisciplinary domain of research now formalized as an annual workshop co-located with various major NLP conferences. However, even within this domain, the parole application poses relatively new challenges. First, existing NLLP benchmark datasets generally source from structured written texts; parole documents are loosely-structured dialogues. Second, information extraction from formal written documents centers around named entities and relation extraction. By contrast, much of the text in the criminal context serves the purpose of surfacing, discussing, and correcting case factors, which are not necessarily relational. This means that understanding parole hearings requires both extractive and abstractive tasks, often across multiple sentences, which is known to be challenging even in more structured settings [Wang et al., 2021].

Through the domain application lens and the machine learning lens as described above, we can understand the dissertation as an application of NLP in service of a deeper understanding of the criminal justice system. The third lens through which to view this dissertation is the connective tissue between the first and the second: *How* can advances in NLP enable social sciences and social

[3]Language models do not necessarily use words as the base unit that input is broken up into [Marcus et al., 1993]. Neural models typically rely on *subword tokenization*, splitting a single word into one or more tokens [Schuster and Nakajima, 2012, Sennrich et al., 2015]. For example, the word "annoyingly" may be read as the two tokens "annoying" and "ly." As such, input length is typically measured in tokens, rather than words.

impact? Our study is one hundred times larger than prior work on California parole hearings, a claim that appeals to proponents of "big data," but what is gained through the increasing scale of research, and by whom? We call this the *integrative lens*.

The present dissertation contributes a new approach for machine learning in criminal justice and builds on a growing body of literature demonstrating the potential for machine learning tools to assist in understanding criminal and administrative justice processes [Lazer et al., 2009, Grimmer and Stewart, 2013, Voigt et al., 2017, Bell et al., 2021]. Machine learning is no replacement for the expertise of a legal scholar, but it can play a complementary role [Abebe et al., 2020]. No algorithm or statistical model can fully inhabit the intricate legal, historical, social, and emotional depth of a parole hearing. But no scholar can reasonably digest a hundred thousand hours of dialogue, spanning over a decade, and glean trends with quantitative precision [Michel et al., 2011, Lieberman et al., 2007].

To lay the groundwork for viewing the rest of the dissertation through the integrative lens, Chapter 2 describes the Recon Approach, a new conceptual and philosophical approach for machine learning in criminal justice. The dominant use for machine learning in criminal justice is as a tool to predict future criminal behavior, which is then often used as a tool for making decisions such as policing [Goel et al., 2016, Barrett, 2017, Ferguson, 2017, Shapiro, 2017, Fryer Jr, 2019] and sentencing [Elek et al., 2015, State vs. Loomis, 2016]. We call such an approach the Predictive Approach. A large effort from the machine learning community has focused on the analysis of existing algorithmic criminal justice tools from the lens of bias and discrimination [Ensign et al., 2018, Friedler et al., 2019, Sánchez-Monedero et al., 2020, Rodolfa et al., 2020]. However, much of this literature comes from the perspective of academic study, taking the existence of tools that use the Predictive Approach as an object for analysis. Few others build on the flaws identified in the Predictive Approach to propose alternatives. Barabas et al. [2018], for example, argue that regression is an incomplete tool for prediction. They argue that the value of a predictive regression is not the outcome, or predicted probabilities, but the significance of the covariates, which should then be used in downstream statistical causal inference [Imbens and Rubin, 2015] to identify interventions.

Such conceptual ideas, as well as academic [Baldus et al., 1990] and legal [McCleskey vs. Kemp, 1987, Baldus, 1995, Blume and Johnson, 2012] work in real-world domains that lead to systemic change, inspire one component of the Recon Approach: *recon*naissance, the task of studying decision-makers and surfacing factors that contribute to ongoing unfairness or arbitrariness. The second component of the Recon Approach is *recon*sideration, which is the goal of using technology in the loop with existing processes for decision review and oversight. To the extent that reconnaissance identifies bias or arbitrariness in the decision-making system, reconsideration is an avenue for actively incorporating those insights into social impact. For example, in a system where individuals who have been denied parole are less likely to have their cases reviewed than those who have been granted parole, a reconsideration tool could increase the chances of finding individuals who were denied

parole, who would actually be suitable candidates.

Chapter 3 provides background on the process through which the California Department of Corrections and Rehabilitation (CDCR) grants and denies parole to individuals serving life sentences. The California prison system is particularly important to study for two main reasons. First, California has an enormous prison and parole-eligible population. CDCR houses the largest number of "lifers," or people serving life sentences with the possibility of parole, in the United States. There are more individuals serving life sentences in California than in the next three states, Texas, Florida, and Georgia, combined [Nellis, 2021]. Second, California is widely considered a model state for parole procedures based on the share of its prison population that is eligible for parole and the comprehensive scope of its review process [Mehta, 2016, Slater, 2020].

Each year, CDCR's Board of Parole Hearings (BPH) schedules thousands of hearings for prisoners who have reached their parole eligibility date. During each hearing, a parole candidate is questioned by a commissioner and a deputy commissioner ("the board"). Commissioners discuss the cases of individual candidates in great detail and close the hearing with a decision on whether to grant or deny parole. In 2019 alone, BPH held 6,061 hearings and granted parole in 1,181 cases. BPH assigns the remaining candidates a period of 3 to 15 years that they must continue to serve before they are eligible to re-appear for a parole hearing. For individuals serving indeterminate sentences, this decision can determine whether they will die in prison.

Chapter 4 describes the data we use in our study of the California parole hearing system. California accords BPH a wide range of discretion and only allows for limited public oversight in its parole decisions. The only public information each hearing produces is a written transcript of the dialogue that is, on average, 150 pages (20,000 words). BPH does not release auxiliary information about the cases in a structured format. Prior studies of parole have thus required researchers to read a small sample of hearing transcripts and meticulously code for analysis variables of interest [Bell, 2019, Friedman and Robinson, 2014, Young et al., 2015, Caldwell, 2016]. As a result, the scope of prior studies has ranged from a total of 109 to 754 transcripts and analyzed between 14 and 21 variables. Through a California Public Records Act (CPRA) request and subsequent court order [Superior Court of California in and for the County of San Francisco, 2020b], we obtained a complete corpus of every digitally available parole hearing transcript for candidates serving indeterminate life sentences in California. The resulting corpus contains 35,105 transcripts and constitutes a complete record of all disclosed[4] hearings from 2007 to 2019. In addition to describing the raw data, Chapter 4 also describes the process we undertook to structure the data, which includes feature selection and manual annotation of said features for a subset of hearing transcripts.

Chapter 4 does not merely serve as an enumeration of data processing operations; it sets the backdrop for the following chapters on natural language processing. Without the data, there would be no machine learning lens. And the data only exists as a result of a sequence, a narrative, of many

[4]CDCR continues to withhold a small number of transcripts, citing confidentiality concerns.

decisions made by imperfect, human researchers. Rather than rely on the notion that machine learning models are trained on so-called "ground truth" data, we acknowledge that all datasets are subjective, by the very nature of their construction and presentation. By presenting the details and decisions behind our dataset construction, we hope not only to make more precise the implications of our results, but also to open an ongoing conversation about the fallibility of datasets.

Chapter 5 continues the conversation about dataset subjectivity and integrity and engages datasets beyond the parole dataset. In this chapter, we investigate label errors that arise naturally from human annotation, inspired by both the annotator validation performed in Chapter 4 and the error analyses of Chapter 6. To this end, we first identify a simple way for large language models to detect naturally occurring label errors. Next, we present a new framework for studying realistic label errors. Existing literature on learning with label errors focuses on either synthetic errors, which are relatively easy to identify, or adversarial errors, which are significantly rarer and harder to identify. We present a method to introduce realistic label errors in order to create new benchmarks for the study of errors that naturally occur as a result of human annotation. The primary contribution of Chapter 5 is most easily viewed through the machine learning lens, as a way for dataset curators to improve their data quality, and as a way for machine learning model designers to more realistically assess model performance in the presence of label errors.

Chapter 6 describes the natural language processing techniques that we developed to extract features from the raw text of the hearing transcripts. The framework considers each feature as a single information extraction task. We primarily tackled two challenges. First, we have relatively few training labels available for each feature. Not only is each hearing time-consuming for SMEs to read, but annotators must also be trained through multiple rounds of validation. Second, the length of the hearing transcripts is challenging not only for annotators but also for state-of-the-art large language models.

Section 6.1 employs data programming [Ratner et al., 2016] to solve the two primary challenges. Instead of annotating parole hearings one at a time, SMEs instead write heuristic *functions*, which can be easily computed over all hearings. The data programming approach addresses the scarcity of training data by trading off quality and quantity: every document is assigned at least one label for each feature, but the label generated by the heuristic may be incorrect.

Section 6.2 finds another strategy for incorporating SME knowledge, inspired by the Open-Domain Question Answering paradigm of a two-step Retriever-Reader model [Chen et al., 2017, Das et al., 2019]. Here, SMEs write heuristics for only the Retriever stage, which we call the Reducer, because of its function in retrieving the most relevant passage from a long parole hearing. Having retrieved a short passage, we can now use relatively sophisticated transformer models for precise question answering over the passage.

Chapter 7 identifies findings about the California parole hearing system through a descriptive regression analysis of the case factors identified in Chapter 4. We compare the effectiveness of three

types of regression: one using only structured data provided by CDCR, one using only manually-annotated data for a small sample of hearings, as described in Chapter 4, and one using only the factors successfully extracted using NLP, but over the entire corpus of hearings, as described in Chapter 6. We find that outcomes are disproportionately impacted by multiple factors outside of the candidate's control, such as which commissioner presides over the hearing and whether the district attorney appears at the hearing.

Chapter 8 uses a combination of statistical and linguistic methods to investigate the role that board-appointed and privately retained attorneys play in parole hearing dialogues and outcomes. We first introduce several linguistic features, and we then build on the factor-based analysis of Chapter 7 to place the impact of linguistic features in context with the factors already introduced in previous chapters. Chapter 8 uncovers disparities in the quality of legal representation that parole candidates receive.

Our results show that circumstantial factors introduce a great amount of arbitrariness into the parole decision process for many candidates. By producing the first comprehensive descriptive analysis of America's largest parole system, we hope to highlight opportunities for parole reform through an integrative modeling approach that combines prediction and explanation [Watts, 2017, Hofman et al., 2021]. Our results suggest that circumstantial factors potentially introduce a great amount of arbitrariness into the parole decision process for many candidates. Our work motivates future studies of causal mechanisms in parole and parole "text as data" [Grimmer and Stewart, 2013]. Our methodology demonstrates that machine learning tools can bring reconnaissance to legal hearing text beyond parole [Bell et al., 2021], and the partnerships we have developed over the course of this project lay the groundwork for context-specific insights into developing methods for reconsideration.

Chapter 2

The Recon Approach

The work described in the dissertation is one of many efforts to use machine learning to serve various efforts in criminal law and criminal justice. As introduced in Chapter 1 as the *integrative* lens, the pursuit of this research means that we concern ourselves not only with the computer science methods of how to perform the research, nor only with the legal domain interest in answering specific questions about parole. We also concern ourselves with the conceptual framework through which we view computer science applications for criminal law. In this chapter, we describe two categories that the majority of such applications fall into, and argue for a complementary approach, which we call the Recon Approach.

First, most applications implicitly work in service of the pursuit of *codified justice*, or the standard application of specifiable rules. Codified justice predates computer "code" and machine learning [Eaglin, 2017, Mayson, 2018], but it is not surprising that it is a notion of fairness that machine learning is well-suited to achieve. The rules need not be explicitly specified in computer code. For example, the parameters of a machine learning model could be interpreted as a set of specifiable rules, so long as they are applied in a standard way across all decisions.

Second, much of the existing technology is *predictive*; it is designed to predict the likelihood that an individual will commit a crime in the future. The intended users of this predictive technology include police officers deciding whom to stop [Goel et al., 2016, Barrett, 2017, Ferguson, 2017, Shapiro, 2017, Fryer Jr, 2019], judges deciding whom to retain in custody pre-trial [Kleinberg et al., 2018a] and what sentence to impose [Elek et al., 2015, State vs. Loomis, 2016], and parole boards deciding whom to keep imprisoned [Reingold and Thomas, 2017]. We broadly categorize this approach as The Predictive Approach.

We argue that the application of machine learning to criminal law does not *necessarily* need to pursue codified justice, nor does it necessarily need to be predictive. In contrast to codified justice, equitable justice is the idea that in order for decisions to be fair, decision-makers need to apply moral principles to unique factual situations and explain their reasoning in doing so. Equitable

justice requires discretionary moral judgment, which facilitates a case-by-case approach.

We propose the Recon Approach, which recognizes the importance of human discretionary judgment in legal decision-making and aims to develop technological tools that provide data-driven opportunities for improving fairness and consistency [Greenawalt, 1975, Hart, 2013]. The Recon Approach is not designed to predict the behavior of defendants, prisoners, and other individuals processed through the criminal legal system. Instead, it is designed to scrutinize how judges, parole board members, and other decision-makers exercise discretion in the context of criminal law. These technological tools operate only in a post-hoc manner. They rely on human beings to make initial judgments and, only after those judgments have been made, find patterns in those decisions and mirror them back. The intended users of the Recon Approach are not frontline decision-makers. Rather, the intended users are the individuals and institutions that investigate decisions.

The Recon Approach consists of two interrelated functions: *reconnaissance* and *reconsideration*. Reconnaissance involves the systematic analysis of a set of decisions to identify what factors tend to influence human decision-making in that context. Reconsideration brings the level of analysis down to individual cases. It involves identifying particular cases that appear to be inconsistent and worthy of a review, or a second look.

Section 2.1 proposes an alternative to the Predictive Approach for decision-making, which is commonly overlooked in machine learning approaches. Section 2.2 further defines codified justice and equitable justice and the relationship between the two. In both these sections, we do not suggest that equitable justice is superior to codified justice, or that predictive approaches are inherently unjust. The goal of the sections is to provide clarifying frameworks to enable a new and complementary type of discourse around decision-making. We hope that a more nuanced understanding of notions of justice can enable individuals to make explicit what a decision-making system or what a machine learning system should value, rather than embedding implicit values and assumptions into a system, and allowing those assumptions to silently propagate.

Section 2.3 defines the Recon Approach, and in particular explains its two components, reconnaissance and reconsideration, and the way that the two components work together. As an orthogonal path of development, the Recon Approach has unique potential that the Predictive Approach is not designed to achieve. Specifically, the Recon Approach aims to protect the role of human discretionary judgment by providing post hoc, data-driven opportunities to improve its fairness and consistency.

The following sections then clarify the niche that the Recon Approach occupies. Section 2.4 argues for the value of discretionary judgment and the role that it plays in legal decision-making. Section 2.5 identifies specific conditions and example applications where the Recon Approach would be particularly relevant. Within those areas where it is most appropriate to apply the Recon Approach, Section 2.6 describes the scope that technological tools can be applied in. In particular, technology is not a panacea or an end-to-end solution; it exists in a fixed scope within an existing

legal system.

The final set of sections address present and future challenges toward the development of the Recon Approach. Section 2.7 sets forth and responds to the most fundamental challenge of the Recon Approach: the concern that it will perpetuate the status quo and its existing inequities. Section 2.8 explains why development of NLP technology is integral to the long-term success of the Recon Approach. Sections 2.9 and 2.10, respectively, discuss the technological challenges and the political challenges which need to be overcome in order to successfully execute the Recon Approach.

2.1 The Predictive Approach

Prediction has become a central value to much of the recent applications of machine learning to criminal justice, leading to what we call The Predictive Approach.

For example, predictive policing tools purport to identify individuals who are more likely to commit crime or geographic areas where crime is more likely to occur [Barrett, 2017]. Police departments in cities like Los Angeles and Chicago have used these tools in deciding to increase preventive policing resources on individuals or areas that the predictive tools have flagged as "hot spots" [Chammah, 2016, Joh, 2017]. In the last five years, seventy percent of police agencies in the United States deployed or increased use of predictive policing technology [Isaac, 2017].

Another common application of the Predictive Approach is an actuarial risk assessment tool purported to estimate the degree of risk that a given individual poses for future violent behavior. Such tools have been developed through analyzing various data sets and identifying correlations between violent behavior and characteristics such as age, prior history of arrests and convictions, employment history, marital status, etc. Algorithms are then developed which take as their input a person's individual characteristics and generate an output indicating the likelihood that a person will commit violence in the future [Starr, 2014]. The basic approach began with statistical models in the 1920s [Tibbitts, 1931, Gross, 2008], but the amount of data considered when generating the algorithms has since increased by orders of magnitude. Given the quantity of data, there is considerable interest in harnessing machine learning to generate improved algorithms [Desmarais and Zottola, 2019, Tonn, 2019]. Currently, criminal law practitioners across the United States use over sixty different risk assessment instruments across various adjudicatory contexts [Barry-Jester et al., 2015, Elek et al., 2015]. Some judges rely on risk assessment scores in making decisions about whether to detain defendants in jail pre-trial and in deciding what sentence to impose upon conviction [State vs. Loomis, 2016, Stevenson, 2018]. In addition, parole board members rely on risk assessment scores in deciding whether to grant people release from prison.

Critics of the Predictive Approach have argued that predictive policing tools and risk assessment instruments are not as accurate as they claim to be [Dressel and Farid, 2018, Tonry, 2019], perpetuate racial bias [Mayson, 2018], and lack adequate transparency [Wexler, 2017, Strandburg, 2019].

Proponents of the Predictive Approach continue working to address these criticisms [Berk, 2019, Bloch-Wehba, 2019, Deeks, 2019]. Proponents also argue that human decision-makers fare no better than algorithms with respect to accuracy, bias, or transparency [Kleinberg et al., 2018c]. In other words, the Predictive Approach may or may not succeed in meeting or surpassing the demands of their critics in terms of accuracy, bias, and transparency.

The ongoing discussion revolves primarily around the ability of the Predictive Approach to achieve its targets defined by accuracy, bias, and transparency in predicting future behavior of individuals. Even if the Predictive Approach does succeed in meeting its goals, legal decision-making systems must also value other goals. Prediction of future dangerousness hasn't always been the aim of criminal justice decisions. In fact, some argue that the shift away from notions of redistribution, reform and rehabilitation, or incapacitation, is itself *driven* by the investment into research and development of the Predictive Approach [Harcourt, 2005].

To use to an example from the educational setting, schools do not generally graduate students based on an assessment of a student's ability to attain employment after graduation. Across the board, whether the graduation decision is based on a simple grade point average or a committee of faculty, the decision to grant a student a degree is based entirely on the existing work the student has completed during the student's program, not any future behavior of the student.

The Predictive Approach does not simply focus on all future behavior; it evaluates, in particular, the behavior of the individuals about whom decisions are made. The availability of data, and the computational resources required to refine the data, has long existed as a power imbalance between decision-makers and those they scrutinize. In intensifying the power imbalance, the Predictive Approach fails to take advantage of the potential of technology to instead hold up a mirror to the decision-makers themselves. We hope that the Recon Approach can serve as such a mirror, casting its gaze on both systemic and individual behavior of the decision-makers.

2.2 Codified and Equitable Justice

In presenting the distinct potential of the Recon Approach, it is helpful to draw upon the distinction between equitable justice and codified justice. This distinction is a theoretical, and rather simplified, distinction: in practice, the two notions of justice overlap. The purpose of the distinction is to better understand the role of human discretionary judgment, in order to better understand how machine learning can interact with such judgment [Davis, 1969].

Codified justice is the standard application of specifiable rules, over a set of facts with a fixed scope. The set of specifiable rules can be thought of as a "legal algorithm" that easily applies to a large number of cases [Gillespie, 2014]. Codified justice also aims to establish the total set of relevant factors in advance, thereby precluding the need for individualized proceedings for discovering other factors or debating the facts of the case [Re and Solow-Niederman, 2019].

Equitable justice, broadly construed, is the idea that in order for decisions to be fair, decision-makers need to apply moral principles to unique factual situations and explain their reasoning in doing so [Tasioulas, 1996]. Under equitable justice, decisions are deemed fair insofar as they are justified on what are taken to be morally legitimate reasons [Bray, 2016]. Therefore, the onus of the decision-maker is to provide a case-specific *explanation* that connects the higher level principles, such as retribution and mercy, to the adherence to or rejection of past patterns of decisions [Nussbaum, 1993, Postema, 2023].

Equitable justice requires discretionary moral judgment, which facilitates a case-by-case approach. Compared to codified justice, which aims to apply a consistent set of rules, equitable justice aims to apply a consistent set of higher level principles. And while codified justice seeks to find general patterns, equitable justice allows for and often promotes the setting aside of those patterns in favor of unique circumstances [Germain, 1518, Smith, 2013]. This understanding of equity accords with modern scholarship that characterizes equity "as a model of decision[-]making that emphasizes case-specific judgment, moral reasoning, discretion, or anti-opportunism" [Jacobs, 2005, Bray, 2016].

To use an example from an educational setting, consider the values that schools weigh in deciding which students to graduate. Primary and secondary schools, for example, may rely on a list of required courses and minimum grade point average to determine which students may graduate. For granting bachelor's degrees, universities may allow for different requirements for different cases, such as for different programs. For example, the grade point average required may be the same across a university, but the required courses for a degree in German literature and a degree in biostatistics may have little to no overlap. In contrast, academic institutions rarely have such simple criteria for granting doctorates. Even within a field, such as biostatistics, it is impossible to specify a set of rules that can apply to every student. For example, there is no one requirement for the number of hours spent on various laboratory tasks. Every student is judged on a case-by-case basis.

Both codified and equitable justice have value in a legal system. Codified justice tends to diminish the vices of discretion like arbitrariness and bias [Davis, 1969] while increasing efficiency and consistency. Equitable justice brings in the virtues of discretion, such as individualized attention to unique case factors and explanations of the reasoning underlying each decision.

2.3 The Recon Approach

Even if the Predictive Approach does succeed in meeting its goals, it is simply not designed to fulfill the distinct objective of the Recon Approach: to recognize the importance of human discretionary judgment and provide opportunities to improve its use in legal decision-making. Technologists are investing in the Predictive Approach and may eventually develop that approach in its most idealized form. The Recon Approach, and by extension human discretion, also deserves this investment.

The reader may immediately wonder: how can technology help us do that? Equitable justice has

long been considered the territory of philosophers and jurists, not computer scientists. And perhaps rightly so. The niche for computer scientists working in law, like data scientists and economists, has thus far been conceived as working in the realm of codified justice to maximize a quantifiable good thing (or to minimize a quantifiable bad thing) [Lehr and Ohm, 2017]. The Predictive Approach aptly fits this established niche by working on cost-effective minimization of (future) criminal behavior. But the aim of the Recon Approach, improving the equitable use of human discretion, is far afield. By definition, its aim is not quantifiable along a single metric. The task cannot be boiled down to a traditional type of maximization (or minimization) problem.

Here, however, computer scientists may help fill a very different niche—the regulation of how people use their discretion. Philosophers and jurists have long been articulating and re-articulating the same problem for equitable justice and discretionary moral judgment. The very feature which makes equitable justice valuable—its human sensitivity to the way that values interact with unique factual scenarios—is also what makes it vulnerable to injustices like inconsistency, bias, and arbitrariness [Davis, 1969]. Paraphrasing Justice Marshall, the power to exercise discretion is also an invitation to discriminate [Furman vs. Georgia, 1972]. This invitation becomes stronger in contexts with a greater number of factors influencing discretionary decisions; it becomes harder to identify which cases were decided for inappropriate reasons. Overall, the legal system struggles to square two values that are in constant tension: the value of treating like cases alike, and the value of treating each case individually.

The traditional approach to navigating this dilemma has been to focus on designing a reliable and fair process by which decisions are made. By ensuring that everyone gets the benefit of that same process, there is a formal sense in which people are receiving equal treatment [Stancil, 2016]. There is also reason to believe that a fairer process improves the likelihood that like cases will receive like outcomes. But although robust procedural protections can reduce unfairness in substantive outcomes, they do not eliminate it [Baldus et al., 1990]. As years of trial and error have shown in the administrative law context, "procedural due process has failed miserably in its mission to rationalize frontline decisionmaking" [Ho, 2017].

Technology can provide an additional process to help reduce unfairness in the outcomes of human decisions. In a framework where human beings make thousands of discretionary decisions based on a set of numerous and broad factors, artificial intelligence (AI) can help detect patterns in the application of those factors. Where it identifies a decision that falls outside this pattern, that decision can be flagged as anomalous. The fact that a particular decision is anomalous does not mean that it was wrong or unfair—but simply that the decision is worth a "second look." A decision that appears anomalous may, upon reconsideration, be judged as a good application of the equitable maxim of judging each case on its own unique facts.[1] Or it may be that the decision is unreasonable upon

[1] As Judge Goodman put it in his defense of judicial discretion at sentencing, "[s]eeming disparity is the result of the fundamental judicial philosophy, to judge each case upon its own facts. It is good to have it. For abstract uniformity we do not need the judicial process. The ipse dixit of the rubber stamp will suffice" [Goodman, 1958].

reconsideration. In addition to reconsidering particular decisions, it is also imperative to consider the patterns in the decision set as a whole. If the patterns turn out to hinge on illicit factors—if, for example, the decisions are found to favor one racial group over another—then there is reason to reconsider the entire system of how the decisions are made.

The Recon Approach takes inspiration from others in the social sciences who analyzed patterns in legal decision-making that were then used by stakeholders as a tool for change [Gelman et al., 2007, Arnold et al., 2018]. An example is the work of David Baldus and others who manually collected information from thousands of records in death penalty cases and analyzed trends among those cases [Baldus et al., 1990]. These researchers found that a death sentence is more likely to be imposed if the victim was White rather than Black; this reconnaissance finding led to decades of impact litigation [Baldus et al., 1990] and statutory reform [McCleskey vs. Kemp, 1987, Blume and Johnson, 2012]. The research also facilitated comparative proportionality review, which calls for reconsideration in a given case if death is excessive when compared to the severity of punishment in cases with similar aggravating and mitigating factors [Baldus, 1995]. This type of research and review, however, has been limited by the incredibly labor-intensive task of pulling data from unstructured text. Machine learning and NLP now offer the possibility of streamlining the process to allow for analysis of much larger sets of decisions and for continually updating those sets as new decisions are made. Instead of investigating a random sample of decisions, the Recon Approach calls for analyzing every decision in a given context and contemporaneously flagging anomalous decisions for reconsideration.

To actualize the Recon Approach, machine learning technologists need to develop a set of tools that we call the Recon Toolkit. We have begun developing these tools for use in the context of parole hearings and see potential for much broader application. The tools that we are developing perform two interrelated functions: reconnaissance and reconsideration.

2.3.1 Reconnaissance

Reconnaissance involves the systematic analysis of a set of decisions to identify what factors tend to influence human decision-making in that context. Reconnaissance tools are designed to review hearing transcripts and other documents related to decisions while using Natural Language Processing (NLP) to create a structured dataset. For example, a tool might take as its input a set of 30,000 parole hearing transcripts and output a spreadsheet that lists fifty data points about each hearing, including information such as the underlying conviction, the amount of time served, the number of rehabilitation programs completed, and whether parole was granted or denied. Reconnaissance tools also take the form of machine learning and statistical analysis techniques that are designed to illuminate patterns in how decision-makers tend to weigh different factors when making decisions. For example, these tools include regression analyses and decision trees that show the branching logic that decision-makers appear to follow when making decisions based on various factors. In these ways,

reconnaissance tools allow the public, legislators, or various stakeholders in the decision-making process to better understand how decisions are being made on the ground. With reconnaissance, the public is better positioned to normatively consider the ways in which a system of decision-making may be working fairly on the whole, or alternatively, may stand in need of structural reform.

2.3.2 Reconsideration

Reconsideration brings the level of analysis down to individual cases. It involves identifying particular cases that appear to be inconsistent with most other decisions in a set of cases with similar specified criteria. The focus of technological development here is on building tools for detecting anomalous cases. An example of a technique for detecting anomalous cases involves identifying groups of "nearest neighbors"—cases that are highly similar with respect to a specified set of case-factors—and ascertaining whether a small fraction of those like cases are not being treated alike. The objective of reconsideration is to create an ongoing and updated list of cases that appear to be anomalous and to provide this list to various types of oversight or review boards. For example, the list may be provided to an agency's administrative review unit, to an independent auditor, or even to attorneys seeking to file appeals. Whoever receives the list would then review each case to assess the decision for potential errors or inconsistencies and recommend (or not) that the decision-makers reconsider a case.

2.3.3 Reconnaissance and Reconsideration Work in Tandem

Although reconnaissance tools are distinct from reconsideration tools, they should be used in tandem. In discussions about our pilot, we have often been asked to consider dropping the reconnaissance function and simply building a reconsideration tool—a "reconsideration-only" tool that does not describe the system as it is but only identifies cases that are outliers. The outliers would be given to the Board (or some other body) for potential reconsideration. Data about which of the decisions are indeed altered by the Board (or some other body) could then be used as additional feedback to continually improve a model for the task of finding decisions that will be altered upon reconsideration. Such a tool might achieve a high "hit rate" for cases worthy of reconsideration, but it would do so in an opaque manner. Absent any reconnaissance, the features that tend to influence initial decisions would remain unknown.

This type of reconsideration-only tool is incompatible with the overarching goal of the Recon Approach because it would tend to perpetuate—rather than ameliorate—existing inequities in the exercise of discretion. It would be trained to enforce the consistency of a system without helping us gain awareness about how the system functions as a whole. To see how, suppose for the purpose of this example that a parole candidate's likelihood of being granted parole is significantly reduced if the candidate is Black. (Prior research has shown that the relationship between race and parole-release is incredibly complex, particularly given that race tends to correlate with several other factors

that influence parole decisions [Huebner and Bynum, 2008, Mechoulan and Sahuguet, 2015, Bradley and Engen, 2016, Young, 2016, Bell, 2019, Greene and Dalke, 2020].) Regardless of whether a reconsideration-only tool used race as a factor in its analysis, it could be less likely to flag the case of the Black parole candidate as an anomaly from the general pattern because, all other things equal, being Black would be more consistent with being denied parole. If fewer cases of Black candidates are flagged as anomalies, then fewer would have their decisions altered, and the reconsideration-only tool would receive less positive feedback for flagging cases of Black candidates. At the same time, the tool would be receiving relatively more positive reinforcement for flagging otherwise alike cases of non-Black candidates. A cycle would thus be perpetuated and become further ingrained, without anyone being the wiser about the underlying problem.

To avoid perpetuating inequities, the Recon Approach insists that reconnaissance must come in tandem with reconsideration. Reconnaissance allows for transparency about how the system functions as whole, as well as more apt use of the reconsideration function. For example, if being Black did reduce the likelihood of being granted parole, stakeholders could push for structural reform going forward that would include a race-sensitive anomaly-detection tool. Such a tool could, for example, review cases of all Black parole candidates and then flag cases for reconsideration if the expected decision would have been different if, all other things equal, the candidate were non-Black. An adjusted tool could also ensure that anomalous cases are identified within racial subgroups and that cases for a particular racial group are reviewed with a frequency that matches this group's demographic representation in prisons.

To be clear, the existence of problematic patterns in the exercise of discretion does not mean that decision-makers are malicious or consciously relying on illicit factors when making their decisions. Patterns might be due to idiosyncratic sensitivities—for example, as previously mentioned, one parole commissioner may have a stronger emotional response to crimes with child victims and be less likely to grant parole in such cases relative to other commissioners. If there are patterns that track racial lines, those patterns might be due to the ubiquitous effects of unconscious bias [Rachlinski et al., 2008]. Another cause for problematic patterns might be due to differentials in the way that cases are presented to parole commissioners. For example, prior research found that the likelihood of parole was lower among parole candidates who were not represented by privately retained attorneys.

The goal of the Recon Approach is not to identify the causal root of problematic patterns or assign blame. Statistical causal inference [Imbens and Rubin, 2015] for the purpose of identifying interventions in the criminal justice system [Barabas et al., 2018] is indeed a possible use for the results identified through reconnaissance, but such causal inference is a separate goal from the scope of the Recon Approach itself. Rather, the goal of the Recon Approach is to make problems clear when they would otherwise remain opaque and to provide opportunities to reconsider the cases of those who, for whatever reason, might have gotten the short end of the stick.

2.4 The Case for Discretion

Given that the primary value of the Recon Approach is providing opportunities to improve human discretionary judgment, it is likely to meet criticism from those who see little value in the role that human discretionary judgment plays in law [Huq, 2020]. Why invest in technology that can improve human discretionary judgment when we could instead invest in technology that could replace human discretionary judgment? There are three reasons why discretion in criminal law should be retained.

First, in certain high stakes decisions, particularly those that determine punishment, respect for human dignity calls for a process in which a person is heard by another human being who can meaningfully consider her situation. Even if the outcome of the decision would be the same as an output from a statistical model, there is value to being heard by "one of us"—another human being. That value has been recognized by jurists [Lockett vs. Ohio, 1978], legal scholars [Mashaw, 1981], psychologists [Tyler, 1990], and those directly impacted by the use of algorithms in criminal law. One man who is on a probation program dictated by an algorithm explained his frustration this way: "I can't explain my situation to a computer . . . But I can sit here and interact with you, and you can see my expressions and what I am going through" [Metz and Satariano, 2020].

Second, discretionary judgment is adept at respecting the multiplicity of values at stake in criminal law. The values at stake in deciding who, whether, and how much to punish have never been boiled down into one determinate and quantifiable aim [Bell, 2017]. The law values public safety as well as proportionality of punishment, fairness in assessing factors that mitigate and aggravate culpability, and capacities for personal growth and change [American Law Institute, 2019]. Human discretion, when functioning well, acts as a way to respect and balance these several (and sometimes competing) values to reach a reasonable judgment [Hart, 2013]. In contrast, insofar as reliance is placed exclusively on predictive technologies like risk assessment tools, only the value of predicting and preventing crime is taken into account. This value would be privileged not necessarily because it is any more important but because it is most easily quantifiable [Harcourt, 2005]. By directing technology toward opportunities to improve discretionary judgment, the Recon Approach is more conducive to respecting the multiplicity of values at stake in criminal law.

Third, those who favor replacing human discretion with algorithmic decision-making often rely on a mistaken assumption about the relative rates of improvement in human discretion as compared to algorithmic decision-making. They tend to argue as follows. Humans have had centuries to improve our ability to exercise discretion, and while there have been improvements, humans are still prone to error, bias, and an inability to truly explain their decisions. Algorithmic decision-making, on the other hand, is in its infancy and quickly improving accuracy, reducing bias, and rendering itself explicable. The rate of improvement in the quality of algorithmic decision-making is assumed to continue exceeding the static rate of improvement of human discretion, and in time, the quality of algorithmic decision-making will eclipse that of human discretion and leave it behind. The assumption of this argument is misguided because the rate of improvement in human discretion

is not static.

The Recon Approach calls for the development of technological tools designed to accelerate improvement in human discretionary decision-making by helping discern systemic issues, explaining how decisions are made, and flagging potentially erroneous decisions for reconsideration. The degree to which the Recon Approach can catalyze improvement in the quality of human decision-making remains an open question. The best way to answer the question is to develop the Recon Toolkit and implement it.

2.5 Applications for the Recon Approach

Our pilot work has applied to the context of parole-release decisions, but the general technique of the Recon Approach can extend to a variety of decision-making contexts that meet the following three criteria. First, the decision at issue must involve the exercise of human discretionary judgment. In decision-making contexts where rote application of rules is preferred over discretionary human judgment, the Recon Approach is not useful. The Recon Approach is committed to the position that discretionary human judgment should be used in at least some contexts in criminal law, but does not itself decide what those contexts are. The aim of the Recon Approach is to provide data-driven opportunities to improve discretion in any context where society has decided discretion ought to be present.

Second, there must be records of the discretionary decision that are available and generally include all information hypothesized to be relevant to the decision [Singer and Caves, 2017].

Third, the decisions need to be made at a slow enough rate to be analyzed. Given that a decision to deny parole is not final until 120 days after the hearing, this window of time allows for the Recon Toolkit to process data from an incoming decision and act on reconsideration before the decision is final. In contrast, consider a police officer's decision to use force on a suspect. Even in the highly unlikely case that an officer made a transcript of his or her reasoning in deciding to use force, time would not allow reconsideration of that decision. Reconnaissance tools could discern patterns in how officers tend to use force [Fryer Jr, 2019] and whether a given instance of the use of force was anomalous after-the-fact. But unlike in the hearing context, officer decisions typically have immediate consequences that cannot be undone.

Given these constraints on scope, we see at least three clear contexts where the Recon Approach could be aptly applied: parole hearings, sentencing hearings, and bail hearings. Researchers may also be able to apply the Recon Approach to prosecutorial charging decisions, but only if prosecutors were to provide some form of transcript that described their thought process for each case. Beyond criminal law, the Recon Approach could apply to civil commitment hearings, child custody termination hearings, and immigration hearings. In the realm of administrative law, particularly within the Social Security Administration, technological tools that scrutinize consistency in decision-making

are emerging [Engstrom and Ho, 2020]. While these tools differ from the Recon tools we are developing in the parole context, there is potential for synergistic development across the disciplines of criminal and administrative law.

2.6 The Scope of the Recon Approach

The Recon Approach starts from a place of acknowledging that human decision-makers have value in our legal system which machine learning cannot replace. It also acknowledges that human decision-makers are imperfect in a number of ways. People are not only prone to make factual errors and oversights, but they are also vulnerable to unconscious (or conscious) biases on the basis of categories like race, class, and gender [Rachlinski et al., 2008]. Human judgment is shaped by idiosyncratic sensitivities. For example, one parole commissioner may have a stronger emotional response to crimes with child victims and be less likely to grant parole in such cases relative to other commissioners. These biases and sensitivities lead to inconsistency in judgments across cases; meaning that not all like cases are treated alike. We see such imperfections in human judgment not as a reason to develop technology to replace human judgment, but as a reason to develop technology that helps bring those imperfections to light and provides stakeholders with data-driven opportunities for improvement.

What stakeholders do with those data-driven opportunities is not up to technologists. On the one hand, a parole board could, for example, use tools like the ones we are developing to identify and reverse hundreds or thousands of decisions denying parole. Researchers could use similar tools to discern whether systemic patterns of racial bias infect certain types of decision-making—in bail, probation, sentencing, jury selection, parole, etc.—and if so, legislatures could use that information to restructure how such decisions are made. On the other hand, seeing the very same evidence, a different parole board could reverse only a handful of decisions, and the legislature could tinker with minor changes in the procedures used for decision-making. Any of these actors could trumpet that they are using cutting-edge technology toward the aim of treating like cases alike. Recon tools, like other technological tools, are a means and not an end in themselves. The means do not themselves ameliorate inequity; they provide opportunities to help people do so.

2.7 Defenses Against Perpetuating Existing Problems with the Status Quo

This section turns to a concern that applies to most AI being developed for the legal field, including both the Predictive Approach and the Recon Approach: that the technology is vulnerable to perpetuating existing problems with the status quo and papering over them with technological

sophistication [Engstrom and Ho, 2020].[2] The concern is particularly acute in the context of application to current criminal law in the United States given the crisis of mass incarceration and widespread inequities in criminal law with respect to race and socioeconomic status.

The concern is that in seeking to reduce inconsistencies within a decision set, the Recon Approach will tend to ossify initial patterns found in a historical decision set. Recall that the first step in building a Recon Toolkit is deciding which factors to lift from the text of the hearing ("the chosen factors"). Based on these chosen factors, reconsideration tools are used to flag anomalous cases for reconsideration. A human then reviews flagged cases and may reconsider the decision. The program then receives feedback as to whether the human changed the decision or not. An initial issue with this kind of feedback loop is that it can perpetuate systemic inequities in decisions. As discussed in Section 2.3.3, it is therefore critical to develop reconnaissance tools that are designed to reveal such inequities.

Even with the reconnaissance tools at work, the feedback loop poses additional concerns. The loop will, in time, lead the program to coalesce or plateau around a subset of factors that are "successful" in resulting in changes to decisions. These factors will be limited to those among the chosen factors; recon tools cannot find anomalies with respect to factors that they have not been trained to pay attention to. Additionally, there may be some chosen factors that have a substantial influence, but only on a very small set of decisions ("super-minority factors"). Because factors like these apply to so few cases, they will be less likely to be reinforced. Factors that apply more broadly will tend to be reinforced and will tend to swallow the super-minority factors. The result is that recon tools will promote consistency among the chosen factors that influence the greatest number of cases, but the tools will be vulnerable to both blind spots and tunnel vision. The blind spots are in the tools' inability to recognize the significance of factors that were not included in initial analysis. And the tunnel vision lies in the tools' tendency to be pulled toward factors that influence large swaths of cases and away from highly nuanced factors impacting very few cases.

To address this vulnerability, we propose that any Recon Toolkit be developed in a way that meets the following three guidelines. First, in initial development, "the chosen factors" should be selected by a process that seeks input from a diverse group of stakeholders. The group should include, at a minimum, decision-makers, people about whom the decisions are made (and their attorneys), prior researchers of that decision-type, legislators, and other representatives of the general public. The stakeholders should be queried as to what factors they think should be included in reconnaissance at the outset. The stakeholders should also be queried on a periodic basis after development of the recon tools because decision norms, as well as perceived knowledge of those norms, may shift over time.

[2]See United States v. Curry, 965 F.3d 313, 353 n.1 (4th Cir. 2020) (Wynn, J., concurring) (expressing concern that "talismanic references to technological terms such as 'big data' and 'machine learning' " may obscure the fact that predictive policing algorithms rely on existing data and so may only reinforce problems in the way policing is done rather than fix them).

Second, the Recon Toolkit should be transparent about what "chosen factors" are included in the model. The tools should be accompanied by a list of factors that were included in its initial development as well as all any factors that were proposed but not included. There should be an explanation for why proposed factors were not included. After development, the list should be updated each time stakeholders are queried. In this way, the public is aware of what the Recon Toolkit is tracking and where potential blind spots may lie.

Third, the tools that flag cases for a second look should be compared periodically to a tool that randomly selects cases for a second look. If more cases from the randomly chosen set of cases are reversed as compared to cases the reconsideration tool flags, the reconsideration tool needs to be adjusted. In other contexts, scholars have suggested this approach as a way to compare the performance of an AI tool relative to a random set of cases that undergo conventional review or "prospective benchmarking" [Engstrom and Ho, 2020].

The dual use objection in California parole hearings. Perhaps the most prominent objection to the Recon Approach is analogous to the "dual use" argument for sentencing [Leins et al., 2020]. While we have developed information aggregation tools for a review use case, what is there to stop someone turning that around and using these exact same features and for a codified justice use case?

In the California parole context, employing technology for a predictive, rule-based system requires legislative parole reform and an overhaul of California's approach to criminal data record keeping. As it is currently constructed, the Board of Parole Hearings operates with great discretion. Parole hearings are based only in part on data that is available before the hearing. For example, parole hearings often discuss mitigating pre-commitment factors such as the living circumstances of an individual at the time that the crime was committed, touching on topics such as childhood abuse, gang membership, or neighborhood crime. These data are often not even available in sentencing transcripts. Even for factors that are available in records before the hearing, such as a candidate's disciplinary conduct in prison, the data often only exists in archived handwritten reports that prison staff aggregate prior to the hearing. The data are read out in semi-structured form for the first time by the commissioner during the hearing. It is therefore not possible to extract a meaningful number of the features that are currently considered for a parole decision in California without first conducting a hearing.

Impact on mass incarceration. The California parole context can also serve as a useful case study for understanding another common objection to perpetuating the status quo. A common question about our work is whether it is possible to use automatically extracted factors for increased review of parole grants, thus increasing the rate at which grants are overturned and contributing to the cycle of mass incarceration. The existing parole review process in California makes additional denials and reversals of grants unlikely. Chapter 3 describes the review process in more detail, but to simplify: after a parole hearing, two parole commissioners make a recommendation to grant or

deny parole. In the next 120 days, the decision is reviewed by the Parole Board. Afterward, the Governor has 30 days to review the decision before it becomes final. In practice, all parole grants are reviewed, but both the Parole Board and the Governor's review unit say that they lack the resources to review many denials. If the decision is a grant, the candidate is released from prison and the outcome is final. However, if the decision is a denial, nothing changes; the parole candidate remains in prison.

So what happens if a prisoner is denied parole, but the decision was in fact inconsistent with the parole decision process? It means there is very limited opportunity to reconsider the case, possibly leaving a prisoner incarcerated much longer than necessary. If an analysis based on features extracted using NLP can identify outlier cases, this is actionable. The Governor may request a review, the Parole Board may advance the date of a hearing, or an appeals attorney may petition a court. On the other hand, there exists no basis on which we should assume that either the Governor or the Parole Board would overturn more hearings when provided with more data about the parole process.

2.8 The Importance of Natural Language Processing for the Recon Approach

In our development of the Recon Approach, we have focused a great deal on building NLP tools to identify and extract information from hearing transcripts. It is worth asking why we would develop new tools when we could instead simply ask decision-makers to record the relevant information as they conduct each hearing. For example, a parole board member could complete a "recon worksheet" during or shortly after the hearing that includes multiple choice questions about the parole candidate's crime, the types of rehabilitation programs completed, the number of years served, and all the other data that an NLP tool might be called upon to extract from a given transcript. The recon team would then use machine learning tools to create models of the collected data and to generate lists of anomalous cases, but the team would no longer need to extract information from transcripts.

Having decision-makers complete such a worksheet would certainly be welcome in the short-term, particularly given the challenges in developing NLP tools for the Recon context, such as the ones discussed in Section 2.9. Scholars have proposed this type of work-around as an alternative or precursor to NLP in other contexts: "a first order solution... would be to standardize inputs" [Engstrom and Ho, 2020]. In the long-term, however, there are four reasons why reliance on decision-makers to complete such a worksheet would be inadequate. These reasons explain why development of NLP tools is integral to the long-term success of the Recon Approach.

First, if a decision-maker has to record particularized information at the time of a hearing, then the required information from past hearings, from before the time information started to be recorded,

would not be available. Decisions made at prior hearings could not be analyzed or potentially included on a list of cases for reconsideration. An NLP tool, however, could analyze prior hearings for which there was a transcript, even before data was collected, and therefore include those hearings in a more complete decision model and generate a more comprehensive list of anomalous cases. The ability to include prior decisions is particularly valuable in contexts such as California where a person denied parole may be incarcerated for up to fifteen years before the next hearing.[3]

The second reason for developing NLP tools is because of the difficulties of creating a definitive list of information to record at the time of the hearing. If a relevant factor is missing from the initial recon worksheet that decision-makers are asked to complete after each hearing, then in order to take the factor into account, someone will have to go back through every hearing transcript to make note of the factor. Doing this task manually is likely cost-prohibitive on a large scale. It is likely that there will be factors that are (or will later become) relevant in the decision-making process that were not included on the initial list and for which no information was recorded. This was our experience in the parole context; at the outset, our discussions with stakeholders led to the selection of factors deemed important to the decision-making process. Unsurprisingly, as the study proceeded, new relevant factors were suggested by various stakeholders or were found to be relevant as we understood the process better. This process seems likely to occur across a variety of decision contexts because of limited knowledge at the outset of a study, improved understanding through research, and changes in decision-making over time. Further, society sometimes shifts its views about how to understand what factors are relevant in decision-making. For example, it used to be uncontroversial to do a study on parole hearings that characterized gender as a binary factor (male or female). There is now growing need to include a nonbinary option. We cannot predict what issues will be on the public's radar in ten years, but we can anticipate that some of those issues are not currently on our radar. The critical advantage of developing an NLP tool to conduct information-extraction is that the tool will be able to efficiently search through all past hearings and extract whatever new pieces of information are needed.

The third reason for urging development of NLP tools is that decision-makers are limited in their ability to accurately record all types of information from a hearing that they are themselves conducting. For example, suppose a parole board commissioner was asked to complete a post-hearing worksheet that asked various questions, including whether the parole board used offensive language during the hearing. It is doubtful that the commissioner would forthrightly answer this question in the affirmative if the commissioner called a parole candidate a "smart ass" during a hearing. Our NLP tool, however, was able to pull out this information from a transcript [Todd et al., 2020].[4] In addition, by putting a decision-maker in the role of recording, and thus to some extent characterizing, the factors that underlie the decision, a degree of objectivity is bound to be lost in translation. For example, the way that a parole board commissioner inputs information on

[3]See California Penal Code §3041.5(4) (West 2016).
[4]See California Board of Parole Hearings, Parole Consideration Hearings 4, 36 (January 2015)

a worksheet may be influenced by that commissioner's ultimate decision about whether to grant or deny parole. We observed a case where, at an earlier hearing, the parole commissioner denied parole and, in articulating the reasons to explain that decision, stated that the candidate contested an underlying aspect of the offense.[5] At a subsequent hearing, a different commissioner granted parole and stated that the candidate was not contesting an underlying aspect of the same offense.[6] Nothing about the candidate's version of the offense changed between the two hearings. It is plausible that the first commissioner had decided to deny parole for some other reason, and that doing so influenced his perspective on whether the candidate was contesting the underlying offense. The advantage of an NLP tool is that it can be trained to extract information about a given hearing in a manner isolated from the final decision of that hearing. To be clear, the claim here is not that the NLP tool will be perfectly objective in extracting information, but that there is reason to believe that it will be more objective than a decision-maker doing the extraction task herself.

The fourth reason for urging the development of NLP tools in the Recon Toolkit is that the technology has the potential to identify factors distinct from the factual information-extraction questions discussed above. These factors can be qualitative and more abstract. The ability to extract such factors could be used as an additional method for identifying anomalous cases for reconsideration in at least two ways. First, an NLP tool could be built to flag hearings that contain linguistic anomalies such as a particularly aggressive questioning style, the use of disrespectful words, or an unusually protracted discussion of the underlying offense. Existing research on detecting linguistic patterns in transcripts from police stops provides good reason to be optimistic about continued development here [Voigt et al., 2017]. Second, recent advances in neural network language models have greatly improved the general performance of NLP, which can be measured simultaneously over a large range of tasks, such as translation, summarization, and language generation. These breakthroughs can be leveraged to help train the AI to identify language that appears strange in its context. An early version of such a tool has been developed; but it needs an individual who is knowledgeable about the parole context to provide feedback on whether the identified cases are indeed anomalies of potential interest or are simply red herrings [Todd et al., 2020]. Once given the feedback, the tool can improve its ability to identify cases of interest. This tool would benefit from continued research in language models, especially in conditional language modeling.

Detection of linguistic anomalies can also work in tandem with the extraction of factual information from transcripts. For example, given the identity of the presiding commissioner of the hearing, a model can be built for the specific speech of one legal actor. This model can be used to identify language anomalies with respect to a given set of decision makers, such as parole commissioners who grant parole at the lowest rates or judges that impose the most severe sentences.

For these four reasons, continued development of NLP is integral to the long-term success of the Recon Approach. As described in the next Part, this development is by no means an easy task and

[5] See California Board of Parole Hearings, Parole Consideration Hearings 121 (February 2016).
[6] See California Board of Parole Hearings, Parole Consideration Hearings 215 (August 2017).

considerable investment is needed to make progress. We hope, however, that the description of the Recon Approach thus far has shown that the investment is well worthwhile.

One ethical question about the development of NLP tools is whether features extracted from hearing dialogue can be used as the input to a risk assessment algorithm before a decision is reached. While constructing a such a risk assessment algorithm is possible in theory, we believe that such an algorithm would be hard to construct and virtually meaningless in the context of parole. Unlike applications to sentencing [Chen et al., 2019, Hu et al., 2018, Zhong et al., 2018], the outcome variable for parole is unclear. Lifer recidivism is extremely low (under 3% in California) and it has not risen even as the parole grant rate has increased from 3% to over 20% in the past two decades [Committee on Revision of the Penal Code, 2020].

2.9 Technological Challenges

In the following sections, we discuss the role that machine learning can play in serving justice that is not oriented around prediction. Unfortunately, the term "prediction" in the technical context of machine learning does not necessarily mean the same thing as it does in common vernacular, and we try to distinguish between the two terms. A prediction, in common vernacular, refers to a statement about a future event, something "not yet seen." In machine learning, models are often described as making predictions, which refers to a *statistical prediction* or *statistical inference*. In this technical context, the prediction or inference about data that is "not yet seen" does not necessarily refer to data that is in the future temporally. A statistical prediction could be a statement about a "not yet seen" individual based on knowledge about the overall population, or vice versa, at a single fixed point in time. For example, a statistical model could predict a person's current height based on their current weight and current age, and make no statement about that person's future height. In the present chapter, we use the term "predict" to refer to the common usage of the term, i.e. referring to statements or decisions about the future.

This section discusses some of the technical challenges for developing the tools that are needed to realize the Recon Approach. For reasons of scope, the discussion is limited to tools that are designed to complete two tasks: (1) extracting information from long-form documents and (2) modeling decisions. For each of these tasks, respectively, we first summarize the basic process, explaining what technical advances need to be made and making suggestions for the near-future direction of research and technological development.

2.9.1 Information Extraction

An information-extraction tool uses NLP to find the answers to queries over a set of long-form documents. An example in the parole context would be answering the following question over 50,000 parole hearing transcripts: "What was the parole candidate's commitment offense?" To

create the information-extraction tool, a set of training data is needed which has picked out the answer to queries across a small subset of documents. The NLP tool is created by learning from this training data and then generalizing to the full set of documents. Curating the training data is a critical step in the process and typically involves employing human annotators (also called coders or labelers in the social science community) to read a subset of documents and answer questions about those documents. The task is time-consuming. For example, annotators for our parole project took an average of forty minutes to answer over 100 queries for each parole hearing transcript. The key advantage of an NLP model is that only a subset of the documents needs to be annotated, and the tool can then learn from those annotations and complete the full set of documents.

Recent advances in building larger and deeper neural networks have led to dramatic performance increases across a range of NLP tasks. But even for these advanced models, the complex information aggregation tasks reconnaissance needs to tackle remain extremely challenging. Current NLP systems must overcome at least three technological challenges in order to tackle the types of information-extraction required for the domains in which the Recon Approach can be used.

First, existing techniques have been applied to short passages of approximately 500 to 1,000 words. These techniques do not scale well to parole hearing transcripts which are approximately 10,000 words.

Second, existing techniques tend to do better when the information to be extracted concerns a specific entity. For example, the tool we are developing can answer the question, "What is the name of the commissioner who is presiding over the hearing?" but struggles to extract an answer for the question, "Was the parole candidate under the influence of narcotics when the underlying offense occurred?" The latter question is challenging because narcotics are discussed in different contexts such as a family history of substance abuse, use before the crime, use while incarcerated after the crime, selling narcotics, etc. The recurrence in different contexts makes it hard to pin down whether a given discussion of narcotics is about the underlying offense or about something else entirely. Existing techniques struggle to extract answers to questions about words that refer to multiple things in different contexts throughout a document.

Third, existing technology struggles to answer questions requiring multiple steps of reasoning. For example, consider the question, "If a parole candidate has been written up for misconduct in prison, what was the date of the last write-up?" To answer this question, natural language processing must find whether there are write-ups for misconduct, find the dates corresponding to each write-up, and then identify the most recent. Requiring the NLP model to hop through multiple relations remains challenging with today's technology [Yang et al., 2018].

To reliably extract information, NLP methods need to be developed to be capable of consuming long text all at once and to incorporate "region isolation" technology that, given a query, can isolate the relevant part of a document. Developing a more sophisticated process for curating training data will also be a requisite step for further progress.

The standard approach for curating training data is to employ human annotators to provide simple answers to queries over a subset of documents. For example, an annotator would simply input "2005" as an answer to the following query: "What was the year of the last write-up for misconduct in prison?" A more thorough approach could prompt annotators to provide additional information to support their answer by highlighting each part of the document that discusses write-ups for misconduct. Another promising idea is to build an interactive annotating process where the machine learning system can continue to ask the annotator for more information on particularly challenging question-answer pairs. For example, the model could ask the annotator if it correctly identified the date of the last write-up in a given transcript. Technologists can make considerable progress by pursuing both human-computer interaction and artificial intelligence efforts to identify the types of annotations required for richer, multimodal tasks.

2.9.2 Decision Modeling

The second type of reconnaissance tool aims to model the decision-making process based on the set of information that has been extracted from the text, statistics from the extraction process,[7] and other data that is not included in the text. Regression analysis is often used to perform this type of task [Rubinfeld, 2000].

Regression analysis has established techniques for measuring important characteristics such as how closely the model fits the relationship between the input factors and the output factor, how probable it is that the patterns found by the model are not the result of mere chance, and the relative weight given to the various input factors.

Despite having well-understood statistical properties, regression analysis has at least two limitations when applied to the recon task of modeling decision-making. First, regression models generally assume that the input factors (independent variables like age, time since the most recent disciplinary write-up, etc.) and the output (a dependent variable like whether parole is granted) are continuous numerical values. For example, the factor of age can be 27, 79, or anything in between, like 46.39. Decision-makers, however, rely on many factors that are categorical rather than continuous. An example of a categorical factor is whether or not a parole candidate was convicted of murder. The standard approach to modeling such categorical factors is to use "dummy variables." For example, a 1 would represent that a candidate was convicted of murder, and a 0 would represent that a candidate was not convicted of murder. However, this approach posits the existence of individuals who are "in between" 0 and 1. But it does not make sense to posit that a person can occupy the space of being "in between" or "somewhat" convicted of murder. As the number of categorical variables grows, this problem magnifies. Consider, for example, the bizarre idea of positing someone who is "in between" a White parole candidate who is diagnosed with schizophrenia, has been convicted of

[7] These statistics should include the measure of the reliability with which the NLP tool extracted the correct answer to its queries. In other words, the decision models should be designed with awareness of the NLP models.

sexual assault, and has done a substance abuse program and a non-White candidate who has no such diagnosis or conviction and has done no substance-abuse program. More sophisticated data encoding techniques have been developed to help regression analysis better account for categorical variables, but limits remain.

Second, regression models are limited in their ability to capture the way that decision-making is intuitively understood. A decision is generally not made in a single step by considering all relevant factors at once. Rather, decision-making tends to involve discrete steps or chains of reasoning. A more appropriate tool for reconnaissance on decision-making help would be one that is designed to model multifactorial judgments. To be clear, such a tool would not purport to capture the actual workings of a decision-maker's own thought process. Rather, it would aim to group cases together based on a shared categorical feature, then form subgroups based on another categorical feature, and then sub-subgroups based on another feature, and so on. In so doing, these types of models use a multi-step process that more intuitively captures our understanding of decision-making.

There are multiple ways of developing such a tool. One example is the nearest neighbors model, which requires stakeholders to define a numerical measure of similarity between different cases. A prototype is illustrated and described in Figure 2.1. Decision trees, modeling data points based on a series of yes-no questions, are another family of models particularly well-suited to modeling decision-making in a multi-step manner. An example of this type of model, as applied to a sample of parole hearing decisions, is shown in Figure 2.2.

This figure illustrates an excerpt of a larger decision tree that was generated from a dataset extracted from a sample of parole transcripts in 2014–2015. In this excerpt, only the top three levels of the tree are shown. The tree reads from the top down. At each step, the algorithm partitions the data into a set of denials and a set of grants as best as possible by setting a threshold on one factor of its choice. The top box asks the first question, "Did the parole candidate receive a risk score of 'low risk' on the psychological risk assessment?" If so, the user would follow the left path down; if not, the right path. The box on the bottom right of the first tree represents all transcripts about a parole candidate with a medium or high psychological risk assessment score who also had more than six years since their last disciplinary write-up. Of these hearings, sixty resulted in a denial and twenty-seven in a grant. The boxes are color coded so that if there are more grants than denials that fit the category, the box is green. Otherwise, the box is red. In theory, the tree could continue extending down, adding more factors and more complexity.

To make decision trees useful for the Recon Approach, additional work is needed in two key areas. First, additional tools are required to better describe how well a given decision tree "fits" the data through measures such as statistical significance and robustness.[8] To see why there is a

[8]Robustness refers to the ability of a statistical model to perform well even if the training data is not perfectly representative—for instance, even if historical parole hearing transcripts do not perfectly represent the possible universe of all parole hearings. This means, for example, that the model should not change too drastically to accommodate the inclusion of an outlier or a transcript that contains an annotation or NLP error.

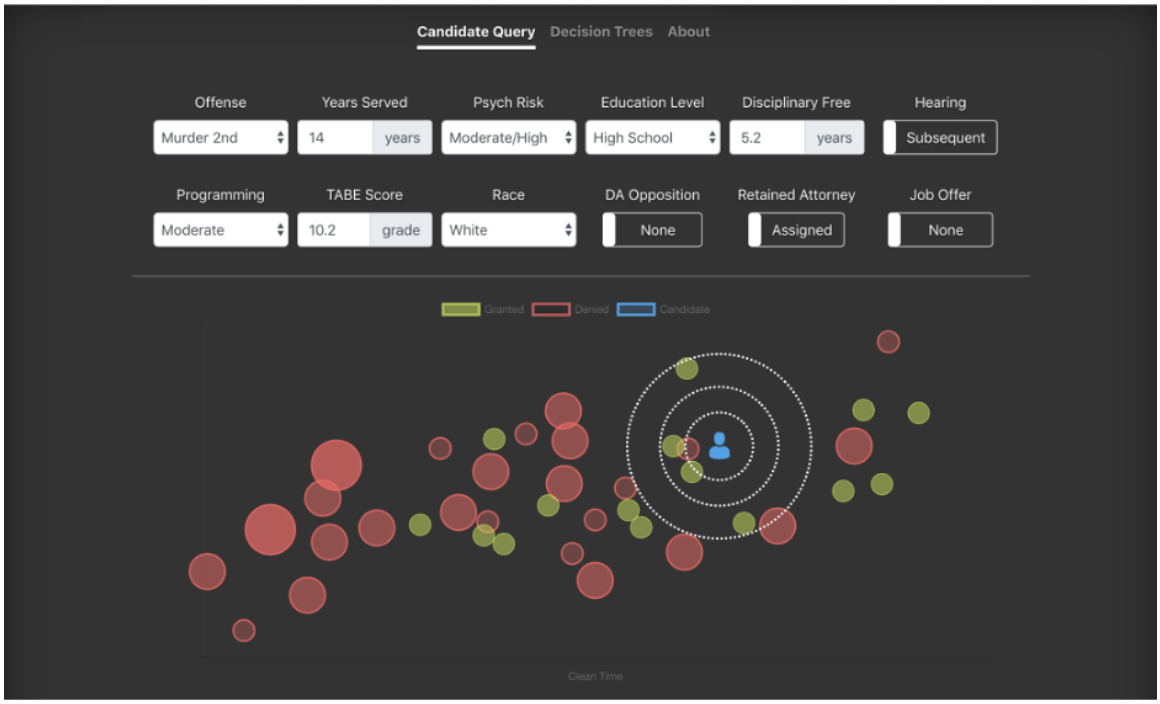

Figure 2.1: This tool shows a stakeholder how an imaginary candidate compares to actual cases that are relatively similar. The stakeholder first inputs information about an imaginary candidate. Here, for example, the imaginary candidate has been convicted of murder in the second-degree, has served 14 years in prison, and so on. Then, that candidate is "plotted" as a blue figure amid actual cases. Green circles illustrate cases where parole was granted, and red circles illustrate cases where parole was denied (the size of the red circle illustrates the period of time that a candidate is scheduled to wait until the next parole hearing – a smaller red circle illustrates a three-year denial period, and a larger red circle illustrates a case with a denial period of five, seven, ten, or fifteen years). The actual cases that are shown on the plot are based on a nearest neighbor calculation. The circles that are closest to the blue figure are most similar to the imaginary candidate. Dotted white rings around the blue figure show which circles would be considered "nearest neighbors" with more restrictive definitions of "near," i.e. only looking at very similar cases.

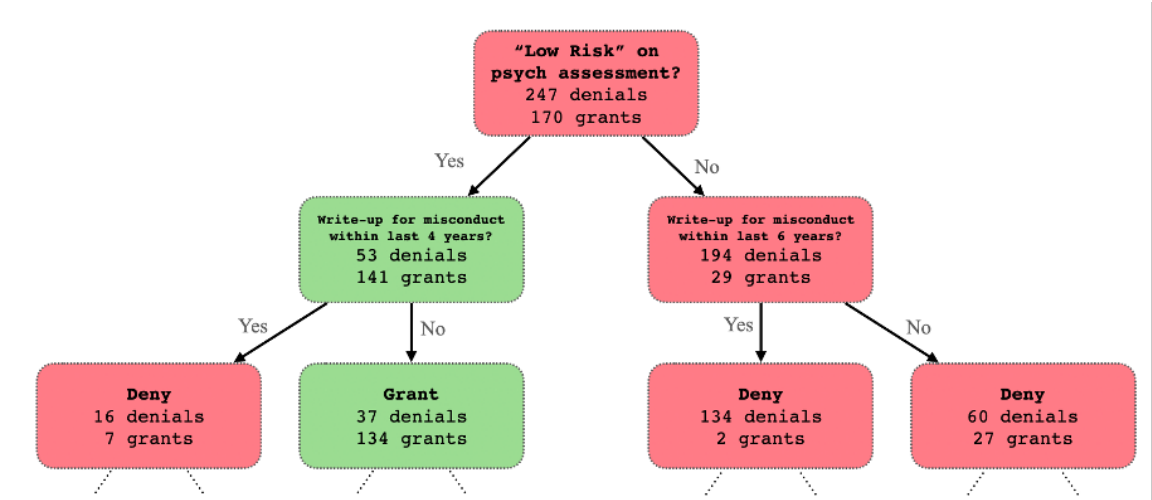

Figure 2.2: Illustrated is an excerpt of a larger decision tree that was generated from a dataset extracted from a sample of parole transcripts in 2014–2015. In this excerpt, only the top three levels of the tree are shown. The tree reads from the top down. At each step, the algorithm partitions the data into a set of denials and a set of grants as best as possible by setting a threshold on one factor of its choice. The top box asks the first question, "Did the parole candidate receive a risk score of 'low risk' on the psychological risk assessment?" If so, follow the left path down, otherwise follow the right path down. The box on the bottom right of the first tree represents all transcripts about a parole candidate with a medium or high psychological risk assessment score, who have also had more than 6 years since their last disciplinary writeup. Of these hearings, 60 resulted in a denial, and 27 in a grant. The boxes are color coded so that if there are more grants than denials that fit the category, the box is green, and otherwise, red. In theory, the tree could continue extending down, adding more factors, and more complexity.

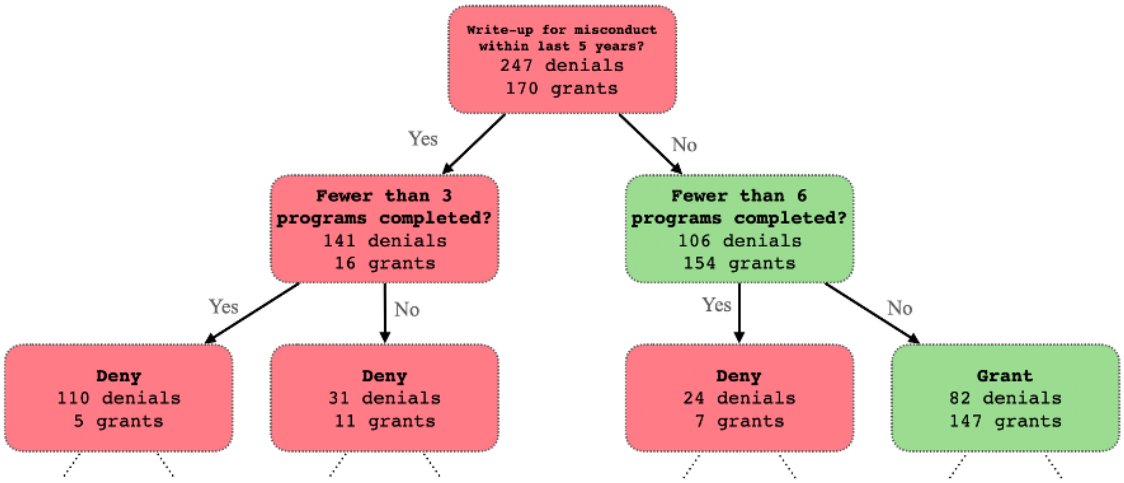

Figure 2.3: Illustrated is an alternative decision tree that was generated over the same set of transcripts as in Figure 2.2. As in Figure 2.2, this is an excerpt of a tree and bottom leaves are not shown.

need for a "fit" metric, consider Figure 2.3 which is built from the same sample of parole hearing decisions as Figure 2.2. It illustrates an alternative decision tree that was generated over the same set of transcripts as Figure 2.2. Again, as in Figure 2.2, this is an excerpt of a tree and bottom leaves are not shown.

The primary criteria for sorting decisions in Figure 2.3 is whether or not a parole candidate received a disciplinary write-up within the last five years. In Figure 2.2, by contrast, the primary criteria are whether or not a parole candidate received a "low risk" score from a psychologist who assessed the candidate prior to the hearing. Each tree seeks to describe the same data, but each was generated by a slightly different algorithm. If one were to take a random set of other cases and follow the chain within the tree, each tree would be roughly equally effective at predicting whether parole would be granted or denied.

What makes one tree a more faithful representation of the pattern of decision-making? In machine learning, this question is largely unexplored. The question that instead receives attention is, "Which tree has a higher degree of accuracy in predicting other decisions?" Techniques have been developed to answer that question, and those techniques have thus far been adequate because trees typically have been used as methods for prediction. For example, multiple trees are often used to form Random Forest models [Ho, 1995] or as part of the XGBoost algorithm [Chen and Guestrin, 2016]. Almost no metrics exist to help choose among multiple trees that predict equally well because the tree's contents do not matter for prediction. Put another way, existing work aims at predicting

which decisions will end up on which decision tree "leaves." The Recon Approach, however, aims to make apt observations about the "branching" within the tree in order to explain the decision-making process.

Additionally, new techniques must be developed to evaluate the quality of the sequencing of the yes-no questions in the tree. How can we know that the branching in a tree like Figure 2.2 more aptly describes a pattern of decisions than Figure 2.3 or some other tree that is generated randomly? Additional techniques are required to answer this question. A model that aptly models decision-making should not be affected by small changes to its input data, such as if one transcript was accidentally omitted or if, for a single hearing, the number of programs completed was incorrectly recorded as "55" instead of "5." Such a model would ideally, for example, not create branches such as, "Did the parole candidate's last name start with the letter P?" A model that goes to great lengths to contort its branches for statistical noise artifacts would most likely not be the most faithful model of the underlying decision-making process—even if such contortions happen to produce correct predictions on historical data.

Decision trees could also benefit from the development of an intuitive way to handle extraction noise. Because the algorithm forming the tree is forced to make a cutoff at each step, it does not easily take extraction noise into account that may be crucial to model. Although social scientists and economists have been modifying regression models easily to handle such noise [Gustafson, 2003], similar methods are lacking for tree-based models. These and other challenges indicate that a substantial amount of future research is needed in order to make the concept of the Recon Approach a practical reality. Our experience thus far has shown that the road ahead is long but well worth pursuing.

2.10 Political Challenges

This section describes two political challenges that the Recon Approach is likely to face and suggests what resources will be needed to overcome these challenges. The discussion is based in large part from experience trying to implement the Recon Approach in the context of parole-suitability decisions in California.

2.10.1 Access to Data

The most pressing obstacle we have faced in implementing the Recon Approach is access to data. Nearly all data about a decision-making process is held by the agency that makes those decisions. The agency has some incentive to resist disclosing data to researchers seeking to implement a Recon Approach: using the Recon Approach may present risks to existing members of the agency. Although the Recon Approach offers a way to improve discretionary decision-making in the long run, it does so by exposing problems with the existing way in which decisions are made. The reconnaissance

process may expose systematic problems in how the agency makes decisions. For example, it may show that, all else equal, a parole board is more likely to give favorable decisions to members of one race relative to another. Additionally, the reconsideration process may expose individual cases that are aberrations from that agency's norm. Bringing public attention to such aberrations can risk tainting the decision-making body's reputation as a whole. Even if there is only one "bad apple," shining a light on it may spoil the whole bunch of decisions in the public eye.

The most promising response to the concern that agencies will deny access to data is ensuring that there is a legal right to access that data. The legal right, however, may be insufficient in practice. For example, our attempts to implement the Recon Approach in the context of the parole board required accessing transcripts of parole hearings as well as relevant information not contained in the transcripts, such as the race of the parole candidates and whether candidates had retained private attorneys for representation at the hearing. Because the transcripts are clearly public records, we were able to obtain them through a public record request. But we were not able to obtain race data because the California Department of Corrections and Rehabilitation (CDCR) withheld it, taking the position that race data was not public record under state law. We postponed our work for approximately nine months of negotiation which led to litigation about our right to access race data.[9] A court held that race data is public record and, in a companion case seeking access to similar data, stated that there is "a weighty public interest in disclosure, i.e., to shed light on whether the parole process is infected by racial or ethnic bias" [Superior Court of California in and for the County of San Francisco, 2020b,a].

Although we were ultimately successful, the time and resources needed for litigation may be cost-prohibitive for many researchers. Furthermore, the uncertainties surrounding litigation and the adversarial nature of litigation can also deter researchers. These litigation costs create an incentive for researchers either to back away from agencies that resist scrutiny or to structure their data requests and data analysis plans in ways that are supportive of, or at least minimally critical of, agencies from whom they are requesting data.

To address this concern, we support efforts to enhance the strength and clarity of public-record laws to make data about decision-making more readily available in practice. Although we successfully litigated in California state court, we would have likely been unsuccessful in a state like Georgia where all information kept by the parole board in performance of their duties is "classified as confidential state secrets."117 Further, we see reason for hope among non-profit organizations like Measures for Justice that have made it their purpose to gather criminal justice data from every county across the country and to make it readily available to the public. We also support development of independent commissions within state governments which are charged to collect and study criminal justice data;

[9]See Verified Petition for Writ of Mandate Ordering Compliance with the California Public Records Act, Voss v. California Department of Corrections and Rehabilitation, No. CPF-20-517117 (Cal. Super. Ct. 2020).

California has recently created such a commission.[10] Lastly, we encourage publication of the "non-finding" that a given agency has refused to disclose data or has restricted access to data after publication of critical findings. In this way, there is at least a small reputational cost that agencies can expect to incur if they deny data to researchers.

In calling for greater public access to decision-making data, we are cognizant of the privacy rights of individuals about whom these decisions are made. We are confident that existing data-security protocols used in other areas of research suffice to protect these rights. For example, in order to begin our research in California, we developed data-security protocols in line with university institutional review boards and California state review board's requirements for human-subjects research.

2.10.2 Researcher Capture

The Recon Approach is potentially vulnerable to a phenomenon that administrative law scholars refer to as "regulatory capture" or "agency capture." The phenomenon occurs when an agency that is charged with independently regulating an industry has had its objectivity compromised by a close relationship with the industry that it is supposed to be regulating. The capture may occur through corrupt means in the form of bribes to the agency from the industry, through more subtle channels such as offering agency-regulators employment opportunities in industry, or through friendships and what has been called cultural capture.

Because the Recon Approach is designed to facilitate oversight over a decision-making body, the researchers implementing the Recon Approach may be liable to capture by the decision-making body itself. As explained above, existing members of the agency have an interest in minimizing the risk that the Recon Approach will uncover problematic issues that could disrupt the regular functioning of the existing agency. This interest may express itself in the form of granting access to only selective data points. It may also express itself in granting access to data only on the condition that any resulting research must be reviewed and approved by the agency prior to publication. Further, a form of capture could occur if researchers are led to believe that their access to data will stop if certain types of criticism are brought into public view. For example, in our efforts to implement the Recon Approach with the Board of Parole Hearings in California, an official asked us to remove from our team a researcher who had published an earlier study finding evidence of racial disparity in the parole process. It was recommended that we replace this individual with the Board's General Counsel—an individual who would represent the Board's interest in making research plans and presenting findings. We declined to do so.

To address this concern, it is important that the agency being studied should not have the power to decide whether or when to withhold data from researchers. In this way, the concern expressed here goes hand-in-hand with the concern expressed above about access to data. Furthermore, institutional

[10] See California Government Code §8286 (West 2019) (creating Committee on the Revision of the Penal Code and requiring that "[a]ll state agencies . . . shall give the commission full information, and reasonable assistance in any matters of research requiring recourse to them, or to data within their knowledge or control.")

review boards that review the ethics of human subjects research ought to review proposals for "capture concerns" when researchers begin a Recon Approach project. Any plan for Recon Approach research should have an explicit commitment to ensuring that research remains independent from influence by the agency that is being studied.

2.11 Conclusion

In his sixteenth-century classic, Utopia, Sir Thomas More wrote, "What you can't put right you must try to make as little wrong as possible. For things will never be perfect, until human beings are perfect—which I don't expect them to be for quite a number of years!" [More, 1551] The Recon Approach can be understood as a technological tool to help answer More's call. The Approach recognizes that, five hundred years later, humans are far from perfect. Its response is not to create a machine to replace human judgment. Such a machine will likewise be imperfect. Instead, the Recon Approach aims to develop tools that act like a flashlight on the past, bringing to light potential problems amid the sprawling web of decisions that humans have already made. In doing so, the Recon Toolkit provides data-driven opportunities "to make [things] as little wrong as possible" Whether those opportunities translate into change is not something we can answer as technologists; it is a question we collectively determine with either action or apathy.

Chapter 3

Background on the California Parole Process

California's Department of Corrections and Rehabilitation (CDCR) houses the largest number of "lifers," or people serving life sentences with the possibility of parole, in the United States. There are more individuals serving life sentences in California than in the next three states, Texas, Florida, and Georgia, combined [Nellis, 2021]. "Lifers" make up approximately 30% of California's 131,000 prison population of 131,000, and as much as half of its current prison population will become eligible for parole consideration during their sentence [Committee on Revision of the Penal Code, 2020].

California is widely considered a model state for parole procedures based on the share of its prison population that is eligible for parole and the comprehensive scope of its review process [Mehta, 2016, Slater, 2020]. Each year, CDCR's Board of Parole Hearings (BPH, or the Board) schedules thousands of hearings for prisoners who have reached their parole eligibility date.[1] In 2019, California scheduled 6,061 parole hearings that resulted in 1,184 grants of parole, a grant rate of 19.5%.

The purpose of each hearing is for the Board to decide whether a given individual who has served enough time to be eligible for release on parole (hereinafter "parole candidate") is suitable for release.[2] State law directs that the Board is to "normally" grant release to parole candidates; the Board is permitted to deny release only if it finds that a candidate "pose[s] an unreasonable risk to public safety". [3]

Parole hearings are generally overseen by one commissioner of the Board and a deputy who

[1] See California Board of Parole Hearings, CY 2019 Suitability Results, `https://www.cdcr.ca.gov/bph/2019/10/24/cy-2019-suitability-results/`.

[2] The Board refers to the hearings as "suitability hearings" and describes the outcome of the hearing as a finding of suitability. For simplicity, we refer to the hearings as "parole-release hearings" or simply "parole hearings" and describe the outcome of the hearing as either granting parole or denying parole. This language has been chosen as more intuitive, but as explained below, a person may be found suitable for parole at the hearing but nevertheless not be granted release if the decision is later reversed.

[3] See California Penal Code §3041(a)(2) (West 2018); *In re* Lawrence, 190 P.3d 535, 560 (California 2008).

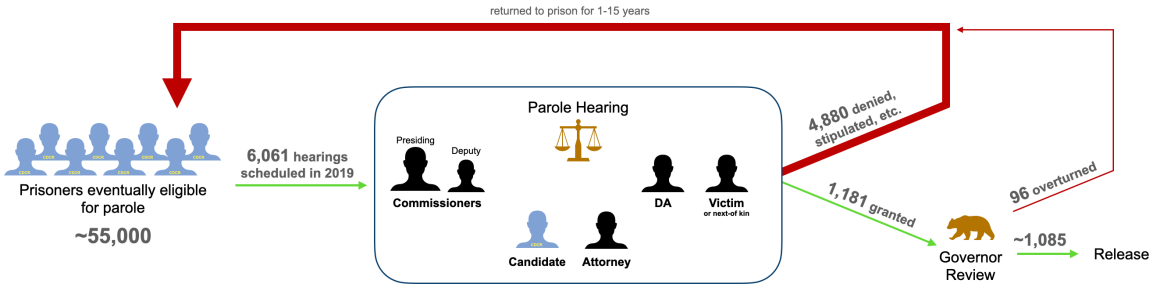

Figure 3.1: An illustration of the parole process in California, annotated with numbers from the 2019 calendar year.

assists the commissioner. The commissioner and deputy ask the parole candidate questions for most of the hearing. The questioning focuses on social history, the underlying crime, the record of conduct in prison, and plans for reentry upon release. This questioning is generally followed by questions and a statement from a district attorney, an attorney representing the parole candidate, and a statement from the victim or victim's next of kin. At the end of the hearing, the commissioner announces whether she finds the parole candidate suitable for release and explains the reasoning for that decision.

If a candidate is found not suitable for release, the commissioner decides whether the next hearing will occur in three, five, seven, ten, or fifteen years.[4] For individuals serving indeterminate sentences, this decision can determine whether they will die in prison.

The Board has broad discretion to decide whether a candidate is suitable for release and must produce publicly available transcripts from each hearing.[5]

The decision made at the hearing is subject to review by the Board's internal administrative review unit as well as California's Governor.[6] The Governor's office has limited resources for decision review; in practice, it reviews all decisions finding parole candidates suitable for parole, but only a small fraction of denials of parole.[7]

Figure 3.1 illustrates the process of parole eligibility, commissioner decision-making, and review by the Governor's office.

If a parole candidate is found unsuitable for parole, the opportunities to reconsider the decision are very limited. A parole candidate can request review by the Board's administrative review unit,[8]

[4]California Penal Code §3041.5 (West 2016).

[5]California Penal Code §3042 (West 2017); *In re* Bode, 88 California Reporter 2d 536, 539 (California Court of Appeals 1999).

[6]See California Penal Code §3041(b)(2) (West 2018) (authorizing the Board to review and reverse decisions); California Constitution Article V, §8 (authorizing the Governor to reverse decisions in murder cases, and to recommend that the Board change its decisions in non-murder cases).

[7]See Interview with staff members who assist Gavin Newsom in review of parole decisions, in Sacramento, CA. (May 13, 2019).

[8]See California Penal Code §3041.5(d) (West 2016) (establishing that parole candidates can petition the Board to advance the date of the next hearing, but petitions are granted only if there is new evidence or a change in circumstances).

as well as judicial review, but there is no right for appointed counsel to do so.[9] On judicial review, the court can vacate a decision by the Board only on the rare occasion that the record contains "no modicum" of evidence that a candidate is currently dangerous.[10] In practice, almost all candidates who are denied parole will remain incarcerated for years until the next opportunity for a parole hearing arises.[11] The wait can last from three years up to fifteen years long.[12]

Although consistency is an aim of parole-release decision-making, it is difficult to measure and achieve given the scale of the system and the Board's breadth of discretion. Short of reading through the hearing transcripts, most of which are 100—150 pages long, there is no readily available data one can analyze to assess the extent to which similar cases receive similar outcomes. The sheer quantity of text makes it nearly impossible to discern whether a parole candidate who is found unsuitable for parole is significantly different from hundreds of others who were found suitable for parole. Further, the fact that administrative regulations direct the Board to consider fifteen factors that are relatively vague makes it difficult to discern what consistency even looks like in this context.[13] For example, one factor that weighs against finding a candidate suitable for parole is whether the offense "demonstrates an exceptionally callous disregard for human suffering."[14] A factor that weighs in favor of finding a candidate suitable for parole is whether "[i]nstitutional activities indicate an enhanced ability to function within the law upon release."[15] Consistency requires treating fittingly similar cases alike, but what makes one parole candidate relevantly like (or unlike) another?

Prior studies of parole-release decisions in California aimed to identify the factors that influence parole decision-making, but the manual labor of reading through hundreds of transcripts limited the sample size of these studies to the range of 100 to 750 parole hearings The sample size limits investigation to a small set of variables, ranging from fourteen to twenty-one variables. Further, given the time required to complete the manual labor of such studies, results have not been released until years after the studied hearings took place [Bell, 2019, Friedman and Robinson, 2014, Young et al., 2015, Young, 2016, Caldwell, 2016]. In the meantime, changes in legislation and administrative regulations make the studies less directly applicable to current decision-making.[16]

[9] *In re* Poole, No. A154517, 2018 WL 3526684, at *14 (Cal. Ct. App. July 23, 2018), reh'g denied (Aug. 21, 2018), review denied (Nov. 14, 2018) ("The role of counsel at the parole suitability hearing is also important because this is the only postconviction stage at which the inmate is entitled to representation by counsel.").

[10] See In re Shaputis II, 265 P.3d 253, 267–68 (Cal. 2011).

[11] See Bell, supra note 17, at 513 (citing Charlie Sarosy, Parole Denial Habeas Corpus Petitions: Why the California Supreme Court Needs to Provide More Clarity on the Scope of Judicial Review, 61 UCLA L. REV. 1134, 1171 (2014)).

[12] See California Penal Code §3041.5 (West 2016).

[13] See California Code of Regulations, title 15, § 2402 (2001).

[14] California Code of Regulations, title 15, § 2402(c)(1)(D) (2001).

[15] California Code of Regulations, title 15, § 2402(d)(9) (2001).

[16] During the time when analysis was ongoing for the studies authored by Friedman and Robinson [2014] and Young [2016], the California legislature passed Senate Bill 260 which changed parole hearings among those under 18 at the time of the offense. See 2013 Cal. Legis. Serv. 312 (West). During the time when analysis was ongoing for the studies authored by Bell [2019] and Caldwell [2016], respectively, the California legislature passed bills that changed parole hearings among those under 26 at the time of the offense, as well as those over age 60 at the time of the hearing. See 2015 Cal. Legis. Serv. 471 (West); 2017 Cal. Legis. Serv. 684 (West); 2017 Cal. Legis. Serv. 676 (West). Between 2015 and 2020, the California Board of Parole Hearings has adopted five different "regulatory packages" that change administrative regulations governing parole hearings. See California Board of Parole Hearings, Recently

3.1 Law Regulating Parole Decisions and Procedures

The legal standard at parole hearings is shaped primarily by statute, and is further informed by administrative regulations and case law. The parole statute provides that the Board "shall normally" grant parole after a parole candidate has served the minimum period of incarceration required by the sentence,[17] unless the Board determines that the candidate "continues to pose an unreasonable risk to public safety."[18] The Board follows administrative regulations that set forth, among other things, lists of reasons that generally support finding a candidate suitable or unsuitable for parole.[19] The California Supreme Court has made clear that while the administrative regulations provide guidance, the ultimate question is whether the parole candidate poses a current danger to the community; if the Board finds that the candidate is not currently dangerous, parole must be granted.[20] The facts of the crime and any pre-conviction history prior to the crime cannot, on their own, support a denial of parole.[21] Such facts can, however, support a denial of parole if there is a "rational nexus" between the crime and current attitudes or recent conduct.[22]

State law provides the following procedural rights: the right to an in-person hearing[23], notice of that hearing, review of the prison file prior to the hearing[24], legal counsel[25], and appointment of legal counsel if a parole candidate is indigent.[26] The Board itself appoints counsel and pays counsel, in contrast to criminal proceedings in which courts appoint public defenders.

There is no right to a hearing in public; hearings take place in prisons where media and members of the public may observe only if a request is approved by the Board.[27] Victims and victims' next of kin have a right to be notified about and attend hearings, but friends, family, or other supporters of the parole candidate have no right to attend hearings and are prohibited from participating in the hearing.[28]

Both the public and the parole candidate have a right to transcripts of hearings.[29] State law requires that the transcripts include everything that is said in the hearing and a definitive, exhaustive statement of the reasons for the parole decision.[30] The transcripts are therefore a reliable source of

Passed Regulatory Packages, `https://www.cdcr.ca.gov/bph/statutes/reg-revisions/` (last visited Apr. 28, 2021)

[17] See California Penal Code §3041 (a)(2) (West, last amended 2017).

[18] See *In re* Lawrence, 190 P.3d 535, 560 (2008).

[19] See California Code of Regulations, title 15, §2402 (2001).

[20] See *In re* Lawrence, 190 P.3d at 554 (current dangerousness is the "overriding" question for the Board).

[21] See *In re* Lawrence, 190 P.3d at 563-64.

[22] See *In re* Shaputis, 190 P.3d at 584-85 (2008).

[23] See California Penal Code §3041.5 (West, last amended 2017).

[24] See California Penal Code §3041.5 (West, last amended 2017).

[25] See California Penal Code §3041.7 (West, last amended 2017).

[26] See California Penal Code §3041.7 (West, last amended 2017); California Code of Regulations, title 15, §2256 (c).

[27] See California Code of Regulations, title 15, §§2029.1, 2030).

[28] See California Code of Regulations, title 15, §2029.1

[29] See California Penal Code §3041.5 (West 2018); *In re* Bode, 88 California Reporter 2d 536, 539 (Court of Appeal 1999).

[30] See *In re* Prather (2010) 234 P.3d 541, 556 (Court of Appeal 2010) (Moreno, J., concurring) ("[T]he Board [is] required to issue a *definitive* written statement of reasons. The Board cannot, after having its parole denial decision reversed, continue to deny parole based on matters that could have been but were not raised in the original hearing.").

both the underlying evidence the Board draws upon and the stated justification for its decisions.

3.2 Record of Evidence at Hearings

The Board considers all relevant and reliable information available in determining parole suitability.[31] Information includes, but is not limited to: records from the underlying conviction; records of misconduct in prison; records of participation in education, vocation, and self-help groups in prison; any essays or self-help book reports that a parole candidate has written; transcripts from prior parole hearings; psychological evaluations (discussed further below); mental health records; written statements by the candidate; letters of support from family, friends and community members; written statements of commendation by prison staff ("laudatory chronos"); documentation of parole plans; letters of opposition; and statements by the victim or the victim's next-of-kin.[32] In some cases, the Board also considers information in the confidential portion of the prison file; this information is not disclosed to anyone at the hearing other than the hearing panel.[33]

In addition, the Board considers a "Comprehensive Risk Assessment" (CRA) report. Shortly before a prisoner's initial parole hearing, a forensic psychologist employed by the Board conducts an interview with the prisoner and writes the CRA report.[34] The psychologist reviews the prison file, which includes, but is not limited to, all the information described above, except for letters of opposition, statements from victims or the victim's next-of-kin, and parole plans if they have not yet been made [Isard, 2017].

3.3 Proceedings at Parole Hearings

Parole hearings are conducted by Board commissioners who the Governor appoints for three-year terms.[35] The Board schedules a parole candidate's first parole hearing approximately one year before the candidate has served the minimum amount of time on the sentence.[36] In many cases, the hearing does not occur on the scheduled date due to waivers, continuances, and postponements.[37] Further, some candidates stipulate that they are not suitable for parole.[38]

[31] See California Code of Regulations, title 15, §2402 (2015).

[32] See California Penal Code §3043 (West, 2016) (referring only to "statements by the victim or the victim's next-of-kin").

[33] See California Penal Code §3042 (West, 2017); California Code of Regulations, title 15, §2235 (2015).

[34] See California Code of Regulations, title 15, §2240 (2015); [California Board of Parole Hearings]

[35] See California Penal Code §5075 (West, 2018).

[36] See California Penal Code §3041 (West, 2018).

[37] See California Code of Regulations, title 15, §2253 (2015).

[38] See California Code of Regulations, title 15, §2253 (2015); [Weisberg et al., 2011]. When a parole candidate waives a hearing, she decides to push the hearing date back one, two, three, four, or five years later. California Code of Regulations, title 15, §2253. When a parole candidate enters a stipulation, she agrees that she is unsuitable for release on parole, and the Board imposes a period of 15, 10, 7, 5, or 3 years until the next hearing. California Code of Regulations, title 15, §2253.

Hearings are conducted in a room inside the prison where the parole candidate is incarcerated. Generally, one commissioner from the Board and one deputy commissioner (the "hearing panel") are present to conduct the hearing and make a finding about whether a person is suitable for release on parole.[39] The attorney representing the parole candidate is present,[40] and a district attorney from the office of the county of conviction may be present in-person or via video conference. Victims and victims' next-of-kin are notified about the hearing in advance; some do not participate, others contribute statements but do not attend, and some attend the hearings in-person.[41]

The vast majority of time at the hearing is devoted to questioning of the parole candidate by the hearing panel. Questions are highly specific to the facts of each case and generally fall into four categories: (i) the candidate's background prior to the conviction, (ii) the underlying offense, (iii) post-conviction activities, and (iv) parole plans. After the questioning period, the district attorney and the parole-candidate's attorney may ask clarifying questions and make closing statements. The parole candidate is then given the opportunity for a closing statement, followed by the victim or the victim's next of kin [Young, 2016].

At the end of the hearing, the Board deliberates and then announces its decision and provides an exhaustive list of reasons for the decision. If parole is denied, the panel determines when the next hearing will be scheduled.[42] The presumptive period of time until the next hearing is fifteen years; the Board may set the time for a shorter period of ten, seven, five, or three years if it finds by clear and convincing evidence that considerations of public safety do not require a longer period of time.[43]

3.4 Decision Review

After the panel present at the hearing makes a decision, the Board's internal Decision Review Unit reviews decisions and may recommend a modification to the decision.[44] The unit may also advance the date of the next parole hearing. If a modification to the decision is recommended, the matter is referred to the full Board, which may rescind or overturn the decision en banc.[45]

Parole decisions are then referred to the Governor who has the authority to reverse the decision in all and only murder cases.[46] In non-murder cases, the Governor is not authorized to reverse parole decisions, but is authorized to review them and request that the Board re-consideration its decision. Governors have varied considerably in the rate at which they reverse Board decisions; for example,

[39] See California Penal Code §3041 (West, 2018).
[40] See California Penal Code §3041.7 (West, 2016).
[41] See California Penal Code §3043 (West, 2016).
[42] See California Penal Code §3041.5 (West, 2016).
[43] See California Penal Code §3041.5 (West, 2016).
[44] See California Code of Regulations, title 15, §2041(h).
[45] See California Code of Regulations, title 15, §2041(h). The Board may reverse or rescind a decision on the basis of an error of law, an error of fact, or new information. See California Code of Regulations, title 15, §2042.
[46] See California Constitution Article V, §8.

Governor Davis reversed 97% of the Board's decisions to grant parole, whereas in 2015, Governor Brown reversed 14% of the Board's decisions to grant parole [Young, 2016].

A candidate who is denied parole may seek review by filing a habeas corpus petition. The court may reverse a decision to deny parole and send it back to the parole board for a new decision,[47] but only if there is nothing in the record which provides "some evidence" of current dangerousness.[48]

3.5 Changes to Law Governing Parole During the Period of Study (2007–2019)

Setback period (2008) Marsy's Law was enacted by California voters in 2008. The law extended the rights of victims at parole hearings and increased the period of time in between a decision to deny parole and the next parole hearing.

Legal standard (2008) Prior to the California Supreme Court's decision in *In re* Lawrence in 2008, parole could be denied solely on the basis of the gravity of the crime. The case established that the overriding question for parole-release decisions is whether the candidate poses a current, unreasonable risk to public safety. The gravity of the underlying crime cannot "in and of itself" provide some evidence of current dangerousness.[49] The gravity of the crime can, however, provide evidence of current dangerousness if some other evidence in the record indicates that the crime remains probative of whether a person is a continuing threat to public safety.[50]

Youth Offender parole (2014) Senate Bill 260 took effect on January 1, 2014 and creates specialized "youth offender parole hearings" for people who committed offenses committed before the age of eighteen.[51] Subsequent amendments to the youth offender parole statute extended eligibility to people serving sentences for crimes committed under age 26.[52] People serving life sentences[53] as

[47] See *In re* Prather, 234 P.3d 541, 544 (2010) (when court grants petition for habeas corpus and reverses Board decision to deny parole, remedy is for Board to conduct a new parole hearing).

[48] See *In re* Shaputis II 265 P.3d 253, 267 (2011) (proper inquiry for court reviewing decision to deny parole is to determine whether whole record "discloses some evidence—a modicum of evidence—supporting the determination that the inmate would pose a danger to the public if released on parole.")

[49] See *In re* Lawrence, 190 P.3d 535, 555 (2008).

[50] See *In re* Lawrence, 190 P.3d 535, 555 (2008). See also *In re* Shaputis I, 190 P.3d 573, 584 (2008) (aggravated nature of crime constituted evidence of current dangerousness to uphold where parole candidate was found to lack insight into a long history of violence).

[51] 2013 California Legislative Service Chapter 312 (West) (amending California Penal Code §§3041 (West, last amended 2017), 3046 (West, last amended 2017), 4801 (West, last amended 2017), and enacting §3051 (West, last amended 2017)).

[52] See 2015 California Legislative Service Chapter 312 (West).

[53] After the legislature enacted penal code section 3051 in 2013, it amended the statute in 2015 and 2017. Under the initial version of section 3051 that was in effect during the time period of the study (Jan. 2014 to June 2015), juveniles sentenced to life without the possibility of parole were ineligible to receive parole hearings under section 3051. In 2017, however, the legislature amended section 3051 to extend eligibility to people who are serving life sentences without the possibility of parole for convictions under the age of eighteen. See 2017 California Legislative Service Chapter 684 (West).

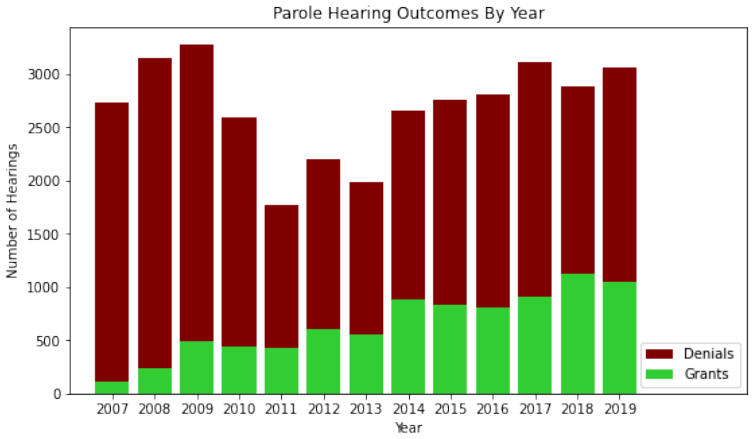

Figure 3.2: Parole hearing grant rates in California 2007–2019.

well as people serving long determinate sentences became eligible for youth offender parole hearings in their fifteenth, twentieth, or twenty-fifth year of incarceration.[54] For some youth offender parole candidates, their initial hearing occurred several years or even several decades earlier than they had anticipated based on the initial sentence. The law also requires that the Board to give "great weight to the diminished culpability of youth as compared to that of adults, the hallmark features of youth, and any subsequent growth and increased maturity of the prisoner in accordance with relevant case law."[55]

Elderly parole (2014) In response to a federal court to reduce the prison population, California began implementing the Elderly Parole Program in 2014. Eligibility extends to individuals who are 60 years or older and who have been incarcerated for at least 25 years. Subsequently, Assembly Bill 3234 extended eligibility to individuals who are 50 years or older and who have been incarcerated for at least 20 years. At an elderly parole hearing, the Board gives special consideration to advanced age, length of confinement, and any diminished physical condition.

[54]See California Penal Code §3051 (a), (b) (West, last amended 2017). The date of the youth offender's initial parole hearing depends on the "controlling offense," defined as the offense or enhancement for which a sentencing court imposed the longest period of incarceration. See California Penal Code §3051(b) (West, last amended 2017). If the controlling offense is a determinate term of years, the youth offender is eligible for release during the fifteenth year of incarceration; if it is a life term less than twenty-five-to-life, the youth offender is eligible for release during the twentieth year; and if it is a life term of twenty-five-to-life or longer, the youth offender is eligible for release during the twenty-fifth year. See California Penal Code §3051(b) (West, last amended 2017).

[55]See California Penal Code §4801(c) (West, last amended 2017).

3.6 Differences in Parole Systems Across the United States

Approximately half a million people are released on parole per year across the United States [Oudekerk and Kaeble, 2021], and it is estimated that release through the discretion of a parole board accounts for one-third to half of all prison releases [Thomas and Reingold, 2017]. The laws, policies, and general functioning of parole boards varies considerably across states [Rhine et al., 2017]. For example, whereas California parole hearings are generally over an hour in duration, parole boards in some other states spend 3–20 minutes per case [Rhine et al., 2017]. In some states, parole-release decisions are made on the basis of paperwork without any hearing or interview with the parole candidate.[56] And whereas California appoints attorneys for parole candidates who cannot afford them, other states prohibit candidates from having legal representation in parole hearings even if they can afford to pay.[57] From 1970 to 2000, the federal government and 15 states largely abolished their parole systems going forward [Renaud, 2019].

Another point of variation is the population who is eligible for parole. In some states, most people in prison are eligible for release on parole; the gravity of their crimes ranges from minor to severe and the length of their maximum sentences may range from two years to life. In California, however, parole eligibility is generally limited to individuals who are serving life with the possibility of parole sentences (subject to exceptions discussed infra). Relative to parole candidates in other states, these individuals have committed crimes of heightened gravity and are far more likely to die in prison if parole is repeatedly denied.

[56] See Vermont Statutes section 502(a).
[57] See *Franciosi v. Michigan Parole Board*, 604 N.W.2d 675 (Michigan 2000).

Chapter 4

California Parole Data

For the remainder of this dissertation, we discuss quantitative methods for analysis of the parole hearing system described in Chapter 3. This chapter describes the process by which we obtain a set of structured features from the hearing transcripts.

Each parole hearing, as described in Section 3.3, is transcribed by a court reporter present at the hearing. The transcription is on the public record, per the California Public Records Act. Through a public records act request accompanied by a research request, we obtained 35,105 transcripts, a complete record of all hearings from 2007 to 2019, from the California Department of Corrections and Rehabilitation (CDCR).[1] Traditionally, this requires compiling and many cycles of editing of an annotation manual, and the training and ongoing supervision of research assistants. We describe this traditional process in Section 4.1.3.

In addition to traditional data annotation, we also pursued two less traditional routes for obtaining structured features about each hearing. First, the manually annotated data is indeed the source of some of the analysis in Chapters 7 and 8, but it is also the basis of training Natural Language Processing (NLP) models, which in turn provide a set of labels, not just for the annotated hearings, but for the entire corpus. This NLP is what enables the scale at which we can analyze the parole hearing corpus. The NLP methods, as they apply to dataset generation, are described briefly in Sections 4.1.10 and 4.1.9. For in-depth explanations and analyses of the NLP methods used, see Chapter 6.

The second less traditional route for obtaining structured features is through a court order. We describe all our data requests to CDCR in Section 4.1.4, and they are best understood not as independent requests, but as an ongoing conversation and efforts to collaborate with CDCR.

In addition to the text of the transcripts themselves, the final products of our data curation efforts are two tables of structured data. The first is a set of 55 features compiled primarily through manual annotation over 688 hearings, and the second is a set of 33 features compiled primarily

[1] CDCR continues to withhold a small number of transcripts, citing confidentiality concerns.

through automated information extraction over 34,994 hearings.

4.1 Methods

The California Parole Dataset accumulated data from a number of channels. The base data is a set of 35,105 PDF transcripts, which is described in the first subsection. We have also taken efforts to obtain a number of structured features associated with each transcript. These features are described in detail in this section; examples include the year of the hearing, the race of the parole candidate, or whether the parole candidate has a job offer.

Each feature is obtained through one or more data sources. We established a fixed ranking of the reliability of our different data sources and extracted each feature from the most reliable source that included the feature for a given transcript. The list, from most reliable to least reliable, is as follows:

1. Manual data annotation ("coding the transcripts") produced by a team of trained annotators. This includes both the 754 transcripts coded for this study as well as 342 additional transcripts coded for other studies for a subset of the features for a total of 1,096 documents.
2. Data obtained through a court order of CDCR in August 2020[2].
3. Data obtained from CDCR through California Public Records Act requests in March 2019.
4. Fields scraped from the CDCR "Inmate Locator" tool.[3] This data source contains limited data about candidates who were granted parole, because their information is removed from the locator upon release.
5. Information parsed from the title pages of the transcripts.
6. Information extracted using Natural Language Processing techniques.

The breakdown of sources for the features included in the analysis is included in Table 4.1. In the following sections, we first describe the raw transcript data and the feature selection process. We then describe each of the processes enumerated above: the CDCR transcript data and manual annotation process (item 1), the organization of CDCR tabular data (items 2–4), and the automated extraction of features from CDCR transcripts using Natural Language Processing (items 5 and 6).

[2]See *Voss v. California Department of Corrections and Rehabilitation*, Verified Petition for Writ of Mandate Ordering Compliance with the California Public Records Act, available at `https://www.courthousenews.com/wp-content/uploads/2020/05/CalifParoleData-COMPLAINT.pdf`

[3]The tool is available online at `https://inmatelocator.cdcr.ca.gov`.

Table 4.1: Origin breakdown for features included in the final dataset spanning all hearings.

Feature	Manual Annotation	CDCR 2020 Suit Results	CDCR 2019 CPRA	Inmate Locator	Title Page Parse	NLP Extraction
`hearing date`	-	100%	-	-	-	-
`candidate attorney`	-	-	-	-	100%	-
`cdcr female`	-	-	-	-	100%	-
`deputy commissioner`	-	-	-	-	100%	-
`presiding commissioner`	-	-	-	-	100%	-
`prison name`	-	-	-	-	100%	-
`da present`	-	-	-	-	100%	-
`retained attorney`	-	100%	-	-	-	-
`ethnicity black`	-	100%	-	-	-	-
`ethnicity latinx`	-	100%	-	-	-	-
`ethnicity white`	-	100%	-	-	-	-
`ethnicity other`	-	100%	-	-	-	-
`proggang`	3%	-	-	-	-	97%
`initial`	-	99%	-	-	-	-
`off mur1`	3%	-	50%	-	-	47%
`off mur2`	3%	-	50%	-	-	47%
`off muratt`	3%	-	50%	-	-	47%
`off sex`	6%	-	94%	-	-	-
`precommit drugsalc`	2%	-	-	-	-	98%
`precommit gang`	3%	-	-	-	-	97%
`tabe`	-	-	-	-	-	100%
`mepd`	2%	-	-	-	-	98%
`victim present`	-	-	-	-	100%	-
`job offer`	-	-	-	-	-	100%
`last writeup`	-	-	-	-	-	100%
`psych assess`	-	-	-	-	-	100%

4.1.1 Transcript Data

Pursuant to California Penal Code 3042(b), the Board of Parole Hearings is required to record and transcribe parole suitability or the setting of a parole date for any prisoner sentenced to a life sentence. The transcription is performed by a court reporter; no audio or video recordings of the hearings are available. Transcripts are considered part of the public record.

We submitted a California Public Records Act request to CDCR on April 11, 2018. On October 1, 2018, the Board of Parole Hearings released all hearings from January 2009 through July 2018. In July 2020, the Board of Parole Hearings released all hearings spanning the 2007 and 2008 calendar years, and all hearings from August 2018 through the December 2019.

The title page of each hearing is well-structured and contains the date of the hearing, the prison at which the hearing is held, the name and CDCR ID of the parole candidate, the name of their attorney, the names of the presiding and deputy commissioners, and a list of others in attendance, though additional participants may or may not be named.

4.1.2 Feature Selection

Factors relevant to understanding the parole process were identified through discussions with legal experts in parole, formerly incarcerated previous parole candidates, advocacy groups including appellate attorneys, representatives from the California Governor's office, and the Parole Board. Discussions with the Board included two conversations with Director Jennifer Shaffer in late 2018 and early 2019. Legal experts in parole have identified relevant features in prior studies: a feature set for a subset of 754 hearings from October 2007–January 2010 [Young et al., 2015] and an adaptation of that feature set for an analysis of a subset of 426 youth offender hearings from January 2014–June 2015 [Bell, 2019]. We included those features identified as more than marginally predictive of binary parole outcome in either of these two prior analysis and added features suggested by the other stakeholders, producing an initial list of 118 proposed features.

We narrowed the initial proposal list down to 36 features in three stages. The first was at the stage of manual data extraction, where we monitored feature reliability at regular checkpoints. If a feature was annotated with very low inter-rater reliability (see Technical Validation), or with very extreme class imbalance, we removed it from our annotation scheme. Some of the proposed features had to be broken down into multiple features; others could be combined into a single feature. The second stage was in post-processing the features after combining data from all sources. If at this stage, additional class imbalance emerges, we discarded the feature. Another reason for discarding features at this stage was for missing data. If one or more data sources provided a feature, but only on a very small subset of documents, this feature was not used.

The final stage was the pre-analysis step. To improve interpretability and further guard against issues arising from class imbalance, we perform further feature combinations at this stage. We test for multicollinearities by calculating Variable Inflation Factors (VIF) and resolve final issues on a

held-out analysis dataset containing approximately 50% of the data used in the full analysis. These three stages yielded the final 36 features. The following sections, and **Feature Refinement** in particular, describe the process of selecting 36 features from the full list of 118.

4.1.3 Manual Annotation

In a study approved by the University of Oregon and the Stanford University Institutional Review Boards (IRBs), a team of 10 paid research assistants was recruited to manually annotate a subset of transcripts. All annotators received IRB training. They were trained over a three-week process during which they all fully annotated the same three transcripts.

We then assigned new transcripts to the research assistants. During the initial rounds of coding, transcripts were triple-coded. For a subset of the triple-coded transcripts, a legal expert who was part of the study team resolved any conflicts during annotator training sessions. After the first month of annotations, two annotators were identified as the most reliable "TAG" annotators. Subsequent transcripts were double-coded (not triple- or quadruple-coded), and the other annotators were only compared to one of the two TAG annotators for the double-coding.

In the first annotation round, 233 transcripts were sampled from the time period between 2007 and 2018, stratified by year. In a next round, a single annotator coded a sample of 500 transcripts from January 2014 to June 2015 for a reduced set of variables. In the third annotation round, 260 transcripts were sampled from the time period between 2014 and 2019, stratified by year, and only the two TAG annotators coded transcripts for the final set of variables.

We proceeded with annotation in multiple rounds using a custom-built data annotation tool depicted in Figure 4.1. On the left side of the tool, a menu shows a list of annotation tasks. Annotation tasks don't necessarily have a one-to-one correspondence with features. For example, a multiple-choice annotation task may be transformed to several binary features. For each annotation task, annotators were asked to extract both the value of the field and click on one or more sentences in the hearing transcript from which they identified the information. An initial set of 118 features proposed as relevant was narrowed down to 59 annotation tasks through subsequent rounds of annotation and reliability checks (some of the proposed fields had to be broken down into multiple tasks, others could be combined into a single task).

4.1.4 Additional Data from CDCR

Accompanying our request for transcript data, we also requested 18 additional features about each hearing. On March 18, 2019, the CDCR Data Requests team provided limited data on the commitment offense and conviction year features for a large subset of 26,760 hearings. CDCR, however, did not release the majority of features that we included in our initial Public Records Act request. In August 2020, we obtained a court order through the San Francisco Superior Court of California for the release of three features: the race of each parole candidate, the current status of each parole

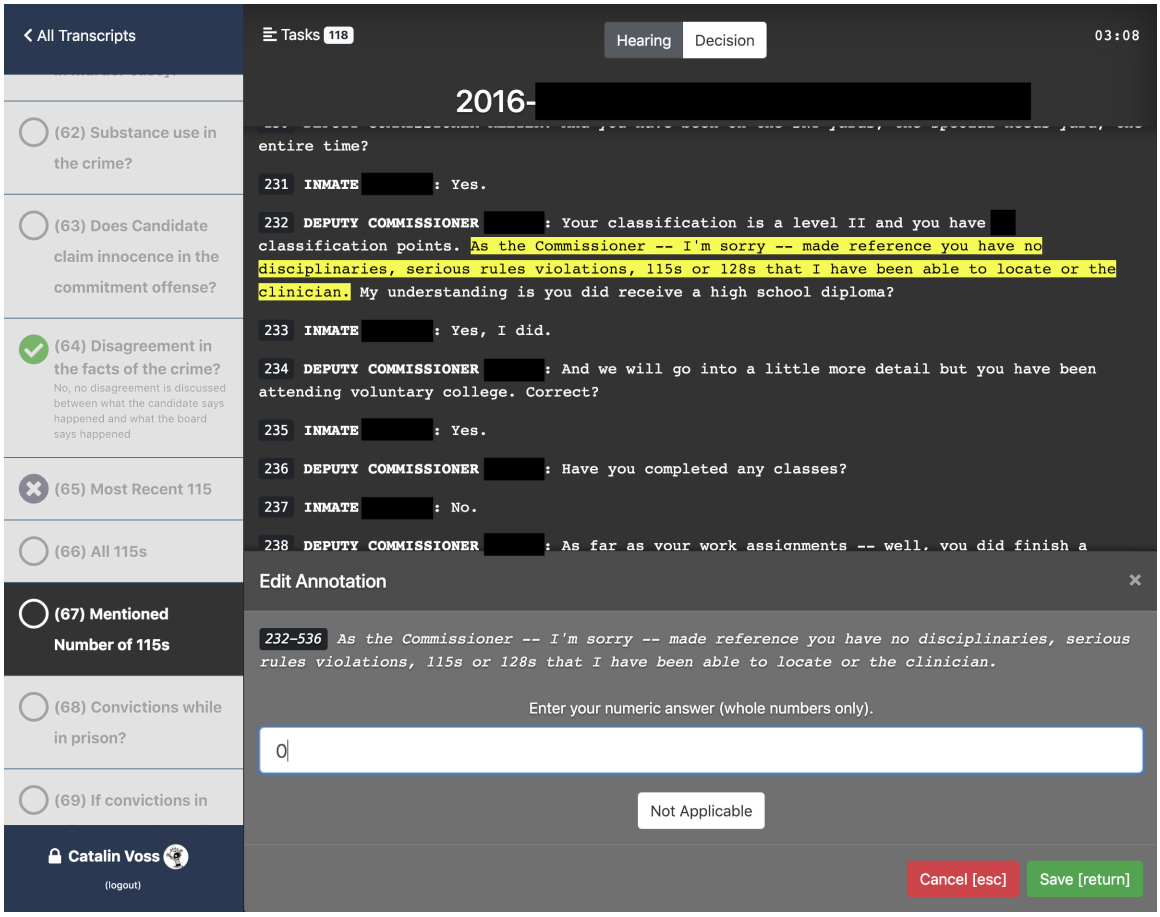

Figure 4.1: A screenshot of the annotation tool used by the annotation team. A timer in the top right corner keeps track of the time annotators spend on each document. In addition to reporting the value for a given annotation tasks, the annotator is asked to click on one or more sentences from which they gleaned the answer.

candidate (released, returned to CDCR, or deceased), and the status of the attorney representing the parole candidate at each hearing (whether the attorney was retained by the parole candidate or appointed by the board). CDCR complied with this court order, providing these three features for all hearings.

As a final source of additional data, separately from the Public Records Act request process, CDCR maintains an Inmate Locator tool on their website that allows the public to look up individuals by name or CDCR ID to find the date they were admitted to CDCR and their parole eligibility date. The tool only contains information about those who are currently incarcerated at the time of the query. We ran the tool on May 10, 2019 and obtained details for 88,645 individuals, of which there were 6,133 who had hearings in our corpus, accounting for a total of 11,748 hearings.

4.1.5 Feature Refinement

A total of 116 features were proposed for analysis. Table 4.4 specifies 22 features that were dropped from the annotation tasks because either a lack of reliability or excessive class imbalance observed early in the annotation process (e.g., nearly all hearings were coded as the same value of the feature). Removing these 22 features, 94 features were carried forward. Table 4.3 specifies 36 additional features that were dropped because of class imbalance observed at the analysis stage. Tables 4.2 specifies the remaining 116-22-36=58 features, which were included for analysis. The final list of features broadly spans the following categories:

- Pre-commitment factors about the candidate's life prior to their conviction
- Factors of the commitment offense
- Post-conviction disciplinary factors describing conduct in prison
- Factors describing participation in rehabilitational programming
- Factors describing post-release parole plans
- Factors of the hearing, such as where it took place and the attorneys and commissioners present
- Factors about the candidate, such as their race and ethnicity
- Whether the candidate retained a private attorney or an attorney was appointed by the board

4.1.6 Feature Transformations

We applied the following transformations to the features, such as transforming or removing features entirely, to deal with mislabeled or missing data. During annotation, annotators were instructed to mark "None" for a feature if the feature was not mentioned during the hearing. Unless otherwise

stated, if a new field is created from a combination of existing fields, then if any of the existing fields is "None," the new field is also "None."

1. All date fields were transformed into years. For the date fields `year received`, `mepd`, and `yped`, we remove all dates before 1920 and treat them as missing data.

2. For all fields corresponding to an offense specified in the California Penal Code (e.g. `off mur1`, `off mur2`), if the field was labeled as "None," we by default set it to be false.

3. For the following fields, we treat a "None" label as the same as a false label, noting that the feature denotes whether mention was present in the hearing: `precommit sexual abuse`, `mental treatment`, `gang debrief`, `prior violence`, `prog12`, `attorney opinion`. We also treat all of `crime drugsalc`, `crime gang`, `crime agent`, and `crime solo` as false if not mentioned unless the candidate was marked as choosing not to discuss the commitment offense.

4. We created a new ordinal variable `gang debrief validate`, which is 2 if `gang validate` is false, 1 if `gang validate` is true but `gang debrief` is also true, and 0 if 1 if `gang validate` is true and `gang debrief` is false. We then discarded the variables `gang validate` and `gang debrief`.

5. We created a new variable `years since last prison convict`, which subtracts `last prison conviction` from `hearing year`, and discarded `last prison conviction`.

6. We created a new variable `prior convictions bucket` which is 1 if `prior convictions` is greater than 5 and 0 otherwise. We then discarded `prior convictions`.

7. We created a new variable `years since eligible` that takes the difference between `hearing year` and the earlier date from `mepd` and `yped`. We then discarded `mepd` and `yped`.

8. We created a new variable `years since received`, which subtracts `year received` from `hearing year`.

9. We created a new variable `justice involved` which is 1 if any of `prior convictions binary` or `prior supervision` or `precommit prison` is true, and 0 otherwise. We then discarded `prior convictions binary`, `prior supervision`, and `precommit prison`.

10. We created a new variable `num prison convictions bucket` which binarizes `num prison convictions`. It is 1 if `num prison convictions` is nonzero, and 0 otherwise. We then discarded `num prison convictions`.

11. We created a new variable `educational bucket`, which is 1 if either `cognitive impair` is true or `tabe` is less than 10. We then discarded `cognitive impair`.

12. For `claim innocence`, we binarized the variable so that `claim innocence` is true if any of the following were true: the candidate claims innocence for all convictions in the commitment offense, the candidate claims innocence for some convictions in the commitment offense, or the candidate chooses not to discuss the commitment offense but does not otherwise specify whether the candidate claims innocence. `claim innocence` is false if the hearing was specifically marked as "no" to the original question, or if it was originally "None."

13. We created a new variable `clean time`, which subtracts the more recent of `last writeup` and `last prison conviction` from `hearing year`. We discarded the variables `last writeup` and `last prison conviction`.

14. We created a new variable `chronos bucket` that is 0 if `chronos` is 0, 1 if `chronos` is between 1 and 9 inclusive, and 2 if `chronos` is at least 10. We discarded the variable `chronos`.

15. We created a new variable `tabe bucket` that buckets the TABE score into 0-8.9, 9-11.9, and 12+ based on histogram analysis. We discarded the variable `tabe`.

16. We created a new variable `prog12 failed` which is true if the candidate was asked about the 12 steps and did not give an adequate response, and is false if either the candidate was not asked about the 12 steps, or if the candidate was asked about the 12 steps and gave an adequate response. We discarded the variable `prog12`.

17. We transform the variable `attorney opinion` so that it is false if the candidate's attorney argued for a setback, and it is true if the candidate's attorney argued for a release, or if it was unclear.

18. We created a new variable `prison is level iv`, which is true if the `prison name` refers to a prison where more than half of the population is level IV, based on `https://www.cdcr.ca.gov/research/compstat/`. We discarded `prison name`. Note that this variable refers to a prison, not an individual. An individual at the time of a hearing is housed at a particular security level, which is not available in the data. Rather, for each prison, CDCR reports the percentage of the prison's overall population that is housed at each security level. Level IV refers to the maximum security level.

19. We created a new variable `years since 2007`, which is the difference between `hearing year` and 2007, the year of the earliest hearing in our dataset.

20. We created a new variable `prog bucket` that is true if the candidate participated in at least four of the following programs: `progang`, `progartfit`, `progedu`, `proggang`, `progparent`, `progphil`, `progrel`, `progsubst`, `progther`, `progvictim`, `progvoc`, and `progoth`. The cutoff is chosen based on a local minimum in the histogram of the sum of the raw programming variables. We discarded the raw programming variables.

21. We created a new variable to summarize the effect of the presiding commissioner for each hearing, rather than assigning as many indicator variables as there are unique presiding commissioners in the dataset. `presiding commissioner rate` is the percentage of hearings the commissioner granted out of total hearings the commissioner presided over prior to the hearing. That is, `presiding commissioner rate` is always calculated using only the facts known at the time of the hearing.

22. Because `year received` and `year convicted` overlap, `year convicted` is dropped, since it has less data available and `year received` is more consistently mentioned in the hearings.

23. Based on histogram inspections, we removed the following features with severe class imbalance ($\geq 95\%$): `confidential info`, `crime agent`, `intimate partner battering`, `gang debrief validate` (95% in the manual sample have never validated), `lifer`, `crime child`, `crime elderly`, `crime police`, and all but three `precommit` variables, `precommit gang`, `precommit sexual abuse`, and `precommit prison`. However, `precommit prison` is included as part of `justice involved` (item 9 above), so was discarded.

24. We discarded all features that are labeled on fewer than 300 documents: `disagree about crime`, `prior convictions bucket`, `residential plan`, `years since last prison convict`.

25. Since the CDCR-provided ethnicity fields constitute a superset of the `race` variable, with the additional "Hispanic"/"Not Hispanic" identification, we only use the ethnicity variables.

To further verify the integrity of our data, we performed the following *sanity checks* on the values of different features. Some of the sanity checks use features that were discarded from the analysis in the previous step. Even though those features may have been discarded from the analysis after being transformed, their raw values are still useful for sanity checks. We enumerate five sanity checks below, and we expect that the statements hold true for all documents. When a statement does not hold true for a given document, we do not necessarily discard the labels for the relevant features for that document. We manually inspected the results. On occasion, we found exceptions to the rules. In other cases, when one of the following statements was not true for a document, it was because of error coming from the data, such as a typo or mistake in human annotation, or an error in NLP extraction. In such cases, we manually corrected the underlying data.

1. We expect the `hearing year` to be at least ten years after `year received`.

2. If `count 115s` is zero, we expect `last writeup` to be "None," and if not, then we expect some value filled in for `last writeup`.

3. We expect `mepd` to be after `year received`. We also expect `mepd` to be *n*o more than one year after `hearing year`.

4. We expect that if `gang validation` is false, then `gang debrief` must also be false.

5. We expect that if `last prison conviction` is specified, then `num prison convictions` should be at least one.

After all these transformations, the 58 features that remained after feature refinement were transformed to 36 features.

4.1.7 Automated Extraction

The following sections describe the automated extraction techniques we used. For the data requests and the manual annotation processes, each method applies to a specific subset of documents (e.g. the 688 sampled for manual annotation, the individuals currently incarcerated for the Inmate Locator tool). However, for automated extraction, each method obtains features for the full set of 34,993 hearings.

4.1.8 Direct Extraction from Title and Closing Pages

Several features are specified on a title or closing page and can be extracted with perfect accuracy. Those features are: the date of the hearing, the prison at which the hearing is held, the name and CDCR ID of the parole candidate, the name of their attorney, the names of the presiding and deputy commissioners, and a list of others in attendance, though additional participants may or may not be named. Although the name of the attorney is listed, the name alone is insufficient for determining whether the attorney is board-appointed or retained, because the same attorney may serve appear in either the board-appointed or the retained role during the course of the timespan of our transcripts. When available, the hearing decision was parsed form the final page of the PDF document. The hearing decision is coded in terms of the number of years a parole candidate is set back before the next hearing. A grant of parole is coded as zero years.

All of the transcripts were parsed using a series of custom PDF parsing tools implemented in Python. To correct for misspelled names of the commissioners and attorneys present in the hearings as well as the name of the prison where the hearing was conducted, we performed clustering was performed using string similarity metrics.

In total, this contributes six features: the hearing date, the prison, the presiding commissioner identity, the parole outcome, and the two binary features of whether a district attorney attended the hearing, and whether a victim or victim's representative attended the hearing. The district attorney's and victim's representative's roles are listed next to the name of each individual, if applicable.

4.1.9 Weakly-Supervised Labeling Functions

Our next approach to automated extraction involves weakly supervised labeling functions, or heuristic functions [Hong et al., 2021b]. We used this method to extract the following nine variables: `proggang`, `off mur1`, `off mur2`, `off muratt`, `precommit drugsalc`, `precommit gang`, `tabe`, `mepd`, and `psych assess`.

Drawing on work in weak supervision, we use the paradigm of data programming [Ratner et al., 2018] to extract some of the factors. We develop a series of "labeling functions" for a subset of the features as a proof of concept. A labeling function is a noisy extractor for a task relying on tools such as regular expressions, string searches, or sentiment analysis. Each set of labeling functions in our implementation includes a prepossessing segmentation function that narrows the text of a long hearing down to one or more smaller chunks of text that are more likely to contain the result. Each labeling function can then either return a result or abstain on the task at the document level.

When combined, multiple labeling functions can comprise a high-quality extractor.[4] We considered several supervised and unsupervised strategies for combining the outputs $\lambda = [\lambda_1, \lambda_2, ...]^T$ from the labeling functions into a single label using limited training data. In our exploratory analysis, we found no benefit from using the unsupervised label aggregation models[Ratner et al., 2018], so we settled on two supervised methods.

- **Logistic Regression** aims to find a set of parameters θ, θ_0 that optimally solves minimizes

$$\min \sum_{x,y} \sigma(\theta^T \lambda(x) + \theta_0) - \textbf{ind.}(y)$$

  where x is a given hearing, y is the true label for this hearing in the training data (as a one-hot vector $\in \mathbb{Z}^K$ for categorical data), and $\sigma_i(z) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$ is the softmax function, for a K-cardinality task. This model is appropriate for situations in which it makes sense to learn a prior θ_0 per category of each tasks. This makes sense for tasks like "Did the candidate participate in gang-related rehabilitational programming?" where it may be reasonable to learn a prior, like that most candidates do participate in such programming. It does not make sense for numerical tasks, tasks with significant class imbalance, or when there is concern of a significant distribution shift between the train and test distribution. For example, for the task of extracting the Minimum Eligible Parole Date (MEPD), there is no point to learning a prior on the "mean MEPD" in the train dataset, and indeed, doing so hurts performance. The logistic regression model can be trained to never abstain as it can utilize its prior to make

[4] The output of the labeling function combination can also be used to train a larger extraction model, such as a neural network with various supervision strategies. We are experimenting with several strategies for doing this, but this remains a challenging task even for large-scale modern language models because of the length of the documents and the task of identifying "where to look." To ensure that the analyses in this study are interpretable, we opted for utilizing only the extractions resulting from combining multiple labeling functions directly and limited our investigation to the tasks that performed the best at this.

a prediction when all of its labeling functions abstain from a prediction.

- Constrained Least Squares (**"linear model"**): To avoid the issues with a prior-based model, we constructed a model which we refer to as the "linear model." This model aims to learn a weighing $\theta \in \mathbb{R}^{|\lambda|}$ of the labeling function outputs λ directly and use this weighting in a majority-rules computation. Intuitively, θ_i tells us how much weight we should attribute to the ith labeling function and thus λ_i. The linear model solves constrained non-negative least squares problem

$$\min \| \sum_{x,y} \max_j \sum_{i=1}^{|\lambda|} \{\theta_i \text{ if } \lambda_i(x) = j \text{ else } 0\} - y\|_2^2$$
$$\text{s.t. } \theta_i > 0 \ \forall \ i \in 1, \ldots, |\lambda|$$

  This model is appropriate for high-cardinality tasks and tasks for which the answer has to be extracted directly from the text and the cardinality cannot be fixed in advance, such as years, numbers or scores.

4.1.10 Pre-Trained Language Models

We used pre-trained language models for the following three variables: `edu level`, `job offer`, and `last writeup`.[Hong et al., 2021a]

For a number of features, the combination of many weak labeling functions fails to extract the correct value with sufficient accuracy. For such features, we leverage advances in neural models for information extraction and question answering. We use an approach inspired by the two-step Retriever-Reader approach to open-domain question answering (ODQA) [Chen et al., 2017, Das et al., 2019]. Here, our two steps are the Reducer and the Producer. We write rule-based Reducers that follow the pattern of generating candidate segments and candidate substrings, [Zhang et al., 2019, Hong et al., 2021a] and sequenced in order of increasing breadth and decreasing precision. The framework provides high-level functions that enable us to easily operate on pipelines of candidate segments, filtering in and out, splitting, de-overlapping, and limiting results to create a high-quality reduced output passage.

Reducers perform the role of segmentation, similar to the preprocessing role in Data Programming. A Reducer selects relevant segments from within a given document. A Producer generates the label from the reduced text. Using a neural model for the Producer provides many advantages in terms of the complexity of text the model is able to digest. However, most neural models are quite limited in the input length of the text it can handle, necessitating a strong Reducer. Many neural models cannot handle more than 500 or one thousand words at a time. Parole hearings are, on average, twenty thousand words, with some much longer than the average. We write and train separate Reducers and Producer for each field of interest.

Using the data produced by the Reducer, we train a Producer for each feature. We use a pre-trained RoBERTa + BigBird (RoB + BB) base model [Zaheer et al., 2020], which is fine-tuned on various prediction heads:

- The feature `job offer` uses a sequence classification head.
- The feature `edu level` uses a sequence classification head.
- The feature `last writeup` uses a masked language modeling head.

4.2 Technical Validation

We describe the validation of the manual annotations in the Manual Annotation section. We use the manual annotations to validate the computer-extracted features.

4.2.1 Inter-Rater Reliability of Manual Annotation

We compute multiple statistics to assess reliability of the annotations. The choice of a reliability metric is not trivial, because the overlap of annotators varies by feature across our dataset and thus Cohen's or Fleiss' κ statistics can only provide incomplete information. We report the following measures:

- Percentage agreement, provided as an uncorrected easily interpretable statistic. For all pairwise documents
- Gwet's AC_1 generalization of kappa for missing data. AC_1 improves over Krippendorff's α statistic which exhibits paradoxical results in the presence of class imbalance due to the formulation of its estimate of expected disagreement. AC_1 has been found to provide a stable inter-rater reliability coefficient with very small bias and mean squared error when the nondifferential error assumption holds [De Raadt et al., 2019].
- Human F1 estimate. In order to provide a point of comparison between our automated feature extraction model evaluation and human extraction performance, we compute a "human F1 estimate." To do this, we consider all pairwise annotations for a feature and designate one of the annotations as gold using the statistical mode, the same logic that is used to resolve labels used for analysis. We then compute the harmonic mean between precision and recall weighted by class prevalence to estimate $\hat{F1}$.

Uncorrected percentage agreement, Gwet's AC_1, and the F1 estimates for all manually annotated features used in the analysis are presented in Table 4.5. For each feature, the answer choices were bucketed into the categories used for analysis before calculating $\hat{\kappa}$. A total of 59 documents was

labeled by at least two raters for all features, and some documents were multi-coded, covering 199 document pairs across 10 annotators.

4.2.2 Limitations of Manual Annotation

The primary purpose of our manual annotation effort is to serve as a baseline following well-established social science methodology, against which we can compare the features extracted from 34,993 transcripts using NLP methods. As such, we did not focus on improving human labeling reliability of the subset of the 118 features initially proposed that proved too difficult to extract for our annotators. Variables for which reliability could not be established were instead cut, resulting in some omitted variable bias. Each one of these 118 factors has a story of its own and many are worthy of much more detailed investigation. Those who wish to pursue research on these features should feel empowered to use the raw transcript text and replicate our annotation methodology, or use a different annotation methodology, for obtaining more annotations.

Because our analysis covers a very large time period with changes in the legislative statues and case law concerning parole, it is not unreasonable to postulate that set of relevant factors changes throughout the analysis period, and that new features may arise over time. For example, legal designations changed over the course of 2007–2019: `confidential info`, `youth offender`, and `elderly parole`.

4.2.3 Evaluation of Extracted Features

During development, we evaluated the labeling functions on a hold-out development set of documents made with annotator- and CDCR-provided annotations. We subsequently trained the combination model on a train set. For each task, we computed accuracy statistic on a validation set and chose the better model. All features with an extraction F1 score of below 0.7 were dropped at this stage. We evaluated the final model on a held-out test set that was never inspected during model development. Table 6.1 lists the dataset sizes for training and validation for each feature.

To increase the amount of data available for training, validation, and testing beyond the data sources described in the present work, we drew on the annotations produced by prior parole studies [Bell, 2019, Young et al., 2015], which were graciously provided by the authors. For the earlier of the two studies [Young et al., 2015], we used the data from only for the validation and test splits to ensure sufficient diversity in our evaluation data.

The mean F1 score for each NLP-extracted feature is described in more depth in Chapter 6. The prior-free linear extraction model outperformed the logistic regression model for all features except for `precommit gang`. This makes sense, since learning a prior for these two binary features likely gives the logistic regression classifier an advantage, but the prior hurts the model for higher-cardinality tasks. The best extraction model for `risk assess` only relies on heuristics. Neural models outperform linear models for three features (`job offer`, `last writeup`, and `edu level`).

Table 4.2: Hypothesized features that are included in the primary analysis, sometimes with transformations or bucketing applied.

Feature	Description
attorney opinion	In the closing statement did the candidate's attorney argue for release, or for a set-back?
claim innocence	Does Candidate claim innocence in the commitment offense?
cognitive impair	Does the candidate have a history of any of the following impairments? (ADHD, Asperger's, Autism, Dyslexia, Low IQ, Designated Speced, Low Cognitive Understanding, Other)
count 115s	Total count of 115s
gender female	Whether candidate female as designated by CDCR number
crime drugsalc	Whether person was intoxicated at time of crime or heavily using alcohol/drugs around the time
crime gang	Was crime rooted in gang activity?
da oppose	DA Opposition
date convicted	Date of last conviction
date received	Date received by CDCR
edu level	Education Level
elderly parole	Elderly Parole Designation - Specific mention of "elderly parole," not based on the candidate's age
eprd	Earliest possible release date - determinate sentences only
ethnicity	CDCR-recorded ethnicity of candidate
gang debrief	Did the candidate previously "debrief" from a prison gang?
gang validation	Validated as gang member?
hearing date	Date of the hearing
initial	Is this the candidate's first hearing?
job offer	Confirmed job offer?
last prison conviction	If convictions in prison, state date of most recent conviction.
last writeup	Most Recent 115
mental illness	History of diagnosed mental illness?
mental treatment	Currently receiving mental health treatment (medication or counseling)?
mepd	Minimum Eligible Parole Date
num prison convictions	Number of convictions while in prison
off mur1	Number of counts of Murder 1 (187)

off mur2	Number of counts of Murder 2 (187)
off muratt	Number of counts of Attempted Murder (664-187 or 217)
off sex	Number of counts of Rape/sexual assault (261-269)
precommit gang	Whether person was involved in gang activity prior to commitment offense
precommit prison	If person was incarcerated prior to commitment offense
precommit sexual abuse	Victim of sexual abuse prior to commitment offense?
presiding commissioner	Presiding Commissioner at the hearing
prior convictions	Number of convictions candidate had before the commitment offense
prior supervision	Select if person was on probation or parole or other form of supervision prior to commitment offense
prior violence	Violent activity prior to the crime? (Includes convictions and admissions)
prison name	Name of the prison where the hearing is being conducted
prog12	Questioned about the 12 steps and whether answered correctly
progang	Anger Programming
progartfit	Art of Phys. Ed. Programming
progedu	Educational programming
proggang	Gang Programming
progoth	Other Programming
progparent	Parenting Programming
progphil	Philanthropic Programming
progrel	Religious Programming
progsubst	Subst. Programming
progther	Cognitive Behavioral or other psychotherapy
progvictim	Victim Programming
progvoc	Vocational Programming
race	CDCR-recorded race of candidate
retained attorney	Whether the candidate privately engaged an attorney
psych assess	Psych Risk Score at most recent comprehensive assessment
tabe	Most recent TABE score (or grade level equivalent if no TABE)
victim oppose	Victim Opposition
victim present	Victim present at hearing?
youth offender	Youth offender parole hearing - 3051/4801/260/261
yped	Youth Offender Parole Eligible Date

Table 4.3: Hypothesized features that were analyzed, but not included in the final analysis due to excessive class imbalance or lack of data.

Feature	Description
candidate attorney	Name of the candidate attorney
age at crime	Age at time of crime (only if stated, not calculated from date)
chronos	How many laudatory chronos did the Candidate receive from members of prison staff?
confidential info[5]	Did the board mentioned that it reviewed confidential information?
current age	Age at the time of the hearing
deputy commissioner	Deputy Commissioner at the hearing
disc contested	whether any discussion about a write-up (115 or 128) being contested
disc sex	whether any discussion of write-ups (115s or 128s) of sexual nature
disc subst	whether any discussion of write-ups (115s or 128s) about substance use/abuse/trade
disc violence	whether any discussion of write-ups (115s or 128s) for violence
lifer	Life with possibility of parole sentence?
off assault firearm	Number of counts of Assault with deadly weapon or firearm - 245(a, b)
off burglary	Number of counts of Burglary (459)
off car theft	Number of counts of car theft and joy riding (VC10851)
off carjacking	Number of counts of Carjacking (215)
off child binary	Was victim a child (under 18)?
off domestic violence	Number of counts of domestic violence (273)
off driveby enhance	Number of counts of driveby enhance -12022.55
off drug possession	Number of counts of drug possession (HS11350)
off elderly binary	If victim was over 65 and/or described as 'elderly' or as 'a senior,' select true.
off false imprisonment	Number of counts of false imprisonment and/or human trafficking (236)
off gang enhance	Number of counts of Gang Enhancement (186.22)
off gun enhance	Number of counts of gun/weapon enhancement - 12021.22-12022.53
off habit enhance	Number of counts of Habitual Offender Enhancement (667, "strike")
off harm to child	Number of counts of harm to child (273)
off injury enhance	Number of counts of great bodily injury enhancement (12022.7)
off invol manslaughter	Number of counts of Involuntary manslaughter (192b)
off kidnapping	Number of counts of Kidnapping (207, 209)
off mayhem	Number of counts of aggravated mayhem (205)
off other counts	Number of other counts
off robbery	Number of counts of Robbery (211, 212)
off shooting at house	Number of counts of shooting at inhabited dwelling (246)
off torture	Number of counts of torture (206)
off vol manslaughter	Number of counts of Voluntary manslaughter (192a)
crime solo	Was crime committed without anyone else helping or pressuring the person into it?
crime police	Was victim a member of law enforcement?

Table 4.4: Hypothesized features that were not included in the analysis at all due to lack of reliability or class imbalance at the annotation stage.

Feature	Description
crime agent	Was the candidate the primary agent that caused harm in the crime [e.g. actual killer in murder case]?
disagree about crime	Disagreement in the facts of the crime?
disc pattern	whether multiple 115s in last five years
intimate partner battering	Victim of intimate partner battering, and committed crime in this context ("battered wife syndrome")
precommit child neglect	Neglect as a child
precommit corporal punish	Harsh corporal punishment (not characterized as child abuse) prior to the crime?
precommit dropout	Dropped out of school (do not mark if left school b/c incarceration)?
precommit drugsalc	Prior drug/alcohol abuse","Whether person had hx of drug/alcohol abuse prior to commitment offense?
precommit drugsalc home	Substance abuse in the home (by someone other than the Candidate) prior to the crime?
precommit emotional abuse	Victim of verbal or emotional abuse prior to the crime?
precommit family victim	Family members or friends victims of violent crime prior to the crime?
precommit foster	Foster care (state number of years in foster care)
precommit homeless	Homelessness prior to the crime?
precommit parent gang	Parent in gang prior to the crime?
precommit parent prison	Parent in prison prior to the crime?
precommit physical abuse	Victim of physical abuse prior to the crime?
precommit suicide	Suicide attempts prior to the crime?
precommit trauma other	Other disadvantage or trauma prior to the crime?
precommit violence home	Witness to violence in home prior to the crime?
precommit violence nonhome	Witness to violence outside of home prior to the crime?
residential plan	Residential plans / confirmed housing arrangements?
yrserved	Total years served (only if stated in hearing, not calculated from date)

Table 4.5: Inter-rater reliability for manually labeled features included in the primary analysis. Three measures are provided: uncorrected percentage agreement, Gwet's AC_1 statistic, which corrects for chance agreement, and a human labeling F1 estimate. A total of 59 documents was labeled by at least two raters for all features, and some documents were multi-coded, covering 199 document pairs across 10 labelers.

Category	**Feature**	Agree	AC_1	$\hat{F1}$
Hearing	victim oppose	1.00	1.00	1.00
	district attny oppose	0.94	0.98	0.96
Pre-Commitment	justice involved *combines features*	0.72	0.82	0.85
	precommit prison	0.76	0.77	0.86
	prior convictions	0.68	0.82	0.85
	prior supervision	0.84	0.85	0.91
	precommit sex abuse	0.97	0.91	0.97
	precommit gang	0.87	0.91	0.93
Commitment Offense	offense murder first	0.93	0.95	0.97
	offense murder second	0.96	0.95	0.98
	offense murder attempt	0.96	0.99	0.98
	offense sex	1.00	1.00	1.00
	crime gang	0.95	0.95	0.97
	crime drugs alcohol	0.75	0.89	0.90
Programs & Rehabilitation	chronos bucket	0.91	0.93	0.91
	programming all *combines features*	0.76	0.95	0.89
	progang	0.75	0.89	0.88
	progartfit	0.82	0.87	0.91
	progedu	0.44	0.56	0.70
	proggang	1.00	1.00	1.00
	progoth	0.53	0.83	0.79
	progparent	0.92	0.91	0.94
	progphil	0.76	0.84	0.91
	progrel	0.78	0.92	0.91
	progsubst	0.93	0.98	0.96
	progther	0.73	0.84	0.88
	progvictim	0.82	0.93	0.93
	progvoc	0.86	0.83	0.87
	12steps program failed	0.94	0.98	0.97
Disciplinary	count 115	0.83	0.95	0.90
	clean time *combines features*	0.76	0.93	0.85
	last writeup	0.72	0.88	0.79
	last prison conviction	0.98	0.95	0.98
	num prison convictions	0.98	0.95	0.98
Parole Preparation	psych assess	0.74	0.83	0.85
	job offer	0.64	0.94	0.83
	years since eligible *combines features*	0.96	0.99	0.97
	yped	0.99	0.90	0.95
	mepd	0.95	0.98	0.96
Special Designation	elderly parole	1.00	1.00	1.00
	youth offender	0.99	0.92	0.98

Chapter 5

Detecting Label Errors by using Pre-Trained Language Models

One of the primary challenges in curating a structured dataset as described in Chapter 4 has been the task of manual annotation and producing an initial set of labels to train NLP models on. However, in spite of many measures of inter-annotator agreement and rounds of annotator calibration, we still identified a small number of errors in the resulting annotations. Some of those errors affected downstream language model training, as described in the error analyses of Chapter 6.

In the process of automating the detection of label errors, we found large, pre-trained language models to be surprisingly effective at this identifying errors in our labels. Simply examining out-of-sample data points in descending order of fine-tuned task loss significantly outperforms more complex error-detection mechanisms proposed in previous work. To validate this method for identifying label errors, we run several additional experiments on commonly-used NLP datasets such as Amazon Reviews and IMDB Movie Reviews.

To this end, we contribute a novel method for introducing realistic, human-originated label noise into existing crowdsourced datasets such as SNLI and TweetNLP. What SNLI and TweetNLP have in common with the parole annotations compiled in Chapter 4 is that their labels are also produced by human annotators, and that as part of the annotation process, a subset of the data is multiply annotated (i.e., multiple annotators assigned to the same data point). The multiple annotations allow us to produce a measure of annotation reliability and to simulate human-originated errors.

We show that the noise of simulated human-originated errors has similar properties to real, hand-verified label errors, and is harder to detect than existing synthetic noise, creating challenges for model robustness. We argue that human-originated noise is a better standard for evaluation than synthetic noise. Finally, we use crowdsourced verification to evaluate the detection of real errors on IMDB, Amazon Reviews, and the parole dataset, and confirm that pre-trained models perform at a

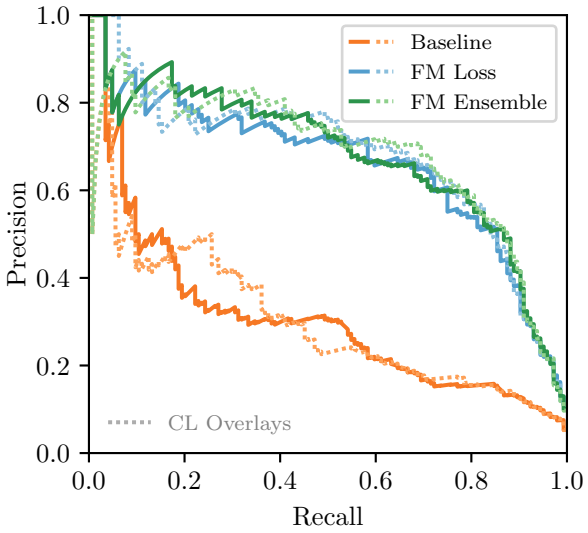

Figure 5.1: Precision-recall curves for label error detection: Large language models detect label errors with high precision, and far more effectively than a baseline word vector-based neural classifier. Overlaying a state-of-the-art model-agnostic error detection method, Confident Learning, results in little to no improvement (TweetNLP-5; §5.7).

9–36% higher absolute Area Under the Precision-Recall Curve than existing models.

5.1 Introduction

Improving model performance in the presence of label errors comprises an area of active research [Song et al., 2022]. However, existing methods focus on label errors in training data. Although seldom acknowledged, evaluation label errors are at least as pernicious as training label errors: pervasive errors in commonly used NLP benchmarks have been found to destabilize model performance [Malik and Bhardwaj, 2011, Northcutt et al., 2021b]. Such findings suggest that improving training methods does not preclude the need for improving the underlying data. We propose a simple method for using large, pre-trained language models (LLMs) to directly identify label errors for the purposes of correcting or removing them.

The majority of work in identifying label errors, and in data-centric artificial intelligence (DCAI) more broadly, focuses on image and healthcare data [DCAI Workshop, 2021]. However, the success of the foundation model (FM) paradigm in applying pre-trained language models to a variety of NLP tasks [Bommasani et al., 2021, Reiss et al., 2020] suggests that FMs may be a powerful tool for detecting and correcting label errors in language datasets. Pre-training has been shown to imbue models with properties such as resistance to label errors, class imbalance [Karthik et al., 2021], out-of-distribution detection [Hendrycks et al., 2018], and confidence calibration [Desai and Durrett, 2020], while conferring robustness, generalization, and natural language understanding capabilities [Wang

Dataset	Text	Label	Sentiment
IMDB	It is really unfortunate that a movie so well produced turns out to be such a disappointment. I thought this was full of (silly) cliches. It had all sorts of differences that it tried to tie together (not a bad thing in itself) but the result is at best awkward, but in fact ridiculous–too many clashes that wouldn't really happen. Then the end of the movie–the last 10 minutes–ruined all the rest. At first I thought Xavier was OK but with retrospect I think he was pretty bad. And that's all really too bad, because technically it was really good, and the soundtrack was great too. So the form was good, but the content pretty horrible.	~~Positive~~	Negative
IMDB	The ending made my heart jump up into my throat. I proceeded to leave the movie theater a little jittery. After all, it was nearly midnight. The movie was better than I expected. I don't know why it didn't last very long in the theaters or make as much money as anticipated. Definitely would recommend.	~~Negative~~	Positive
Amazon	The new design only has a thin layer of cellulose sponge material. It will not last as long. Already showing signs of wearing out. The picture does not represent the item received.	~~Neutral~~	Negative

Table 5.1: Organic label errors from sentiment datasets IMDB and Amazon, shown with the original dataset label. Each example was hypothesized by our model to be erroneous, and later verified by crowd workers.

et al., 2018, Petroni et al., 2019]. Our primary contribution is to show that simply verifying items in order of their out-of-sample loss on a foundation model improves precision by an absolute 15–28% and Area Under the Precision-Recall Curve (AUPR) by an absolute 9–36%.

Many methods for label error detection rely on artificially introduced label errors as ground truth for evaluating their methods. [Northcutt et al., 2021a] develop a state-of-the-art model for identifying label errors, Confident Learning (CL), and use the better approach of crowdsourced human evaluations to determine the ground truth of label errors. We model our experiments on real data after their verification protocol, replicating this on real errors in IMDB [Maas et al., 2011], Amazon Reviews [McAuley et al., 2015], and Recon [Hong et al., 2021a], with adaptations to mitigate annotator fraud [Kennedy et al., 2020].

In the process of assessing our results, we contribute a novel technique and protocol for introducing realistic, human-originated label noise into existing crowdsourced datasets, and apply it to two such datasets, TweetNLP [Gimpel et al., 2010] and SNLI [Bowman et al., 2015]. We demonstrate that our technique better approximates *organic* (real, naturally occurring) label errors than existing methods. We provide evidence that this realism is essential to properly assessing model performance: even models that are robust to standard synthetic noising approaches show limited robustness to human-originated noise.[1]

[1]Data noising library and evaluation data available at https://github.com/dcx/lnlfm.

5.2 Related Work

Learning with Noisy Labels (LNL) focuses on the model-training stage. Noise-robust approaches examine model enhancements such as the design of loss functions [Joulin et al., 2016, Amid et al., 2019, Liu and Guo, 2020, Ma et al., 2020], regularization [Azadi et al., 2015, Zhou and Chen, 2021], reweighting [Bar et al., 2021, Kumar and Amid, 2021], hard negative mining and contrastive learning [Zhang and Stratos, 2021]. Noise-cleansing approaches aim to segregate clean data from noisy data in training, e.g. bagging and boosting [Wheway, 2000, Sluban et al., 2014], k-nearest neighbors [Delany et al., 2012], outlier detection [Gamberger et al., 2000, Thongkam et al., 2008], bootstrapping [Reed et al., 2014], and neural networks supervised directly on detecting an error, when such data exist [Jiang et al., 2018].

LNL methods have in most cases been evaluated using artificially-generated label noise. A typical evaluation of an LNL method uses a standard benchmark dataset, and programmatically corrupts training labels via one of three main noising schemes [Frenay and Verleysen, 2014, Algan and Ulusoy, 2020]. *Uniform noise* is most commonly used but unrealistic; deep neural networks have been found to perform well even when noised labels outnumber original labels at a ratio of 100 to 1 [Rolnick et al., 2017]. *Class-dependent noise* randomly permutes labels based on a confusion matrix. However, research on annotator disagreement suggests that label errors tend to result from feature-based, not class-based ambiguity [Hendrycks et al., 2018]. Training models to generate realistic *feature-based* or *instance-dependent noise* has recently emerged as an area of active research [Chen et al., 2021b, Xu et al., 2021a, Dawson and Polikar, 2021]. However, [Algan and Ulusoy, 2020] report that feature-dependent noise may bias benchmark performance toward similar models to the ones used to generate this noise.

The noising schemes above each fail in some way to simulate organic, naturally occurring label errors, which are estimated to occur in common benchmarks at 1–5% of labels [Redman, 1998, Müller and Markert, 2019, Northcutt et al., 2021b, Kreutzer et al., 2022] or even as much as 20% [Hovy et al., 2014, Abedjan et al., 2016]. For organic errors, CL [Northcutt et al., 2021a] predicts errors in IMDB, Amazon Reviews, and other datasets by estimating a joint distribution between noisy and uncorrupted labels; Reiss et al. [2020] pioneers using BERT for error detection on ConLL-2003 via a classifier trained over a frozen BERT embeddings layer.

5.3 Methods

Motivation. Empirical evidence on image data suggests that models exhibit high loss on label errors in training data relative to the underlying features [Huang et al., 2019, Kim et al., 2021, Hong et al., 2021a, Chen et al., 2021a]. Hendrycks and Gimpel [2017] show that predicted probabilities of (non pre-trained) neural networks can identify out-of-distribution examples. We consider the framing that label errors are one type of out-of-distribution data. Indeed, CL [Northcutt et al., 2021a] uses

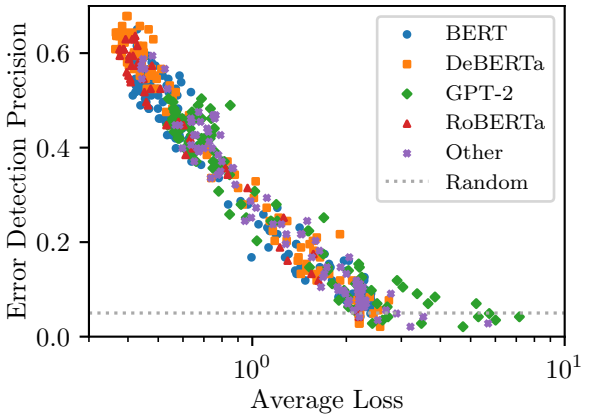

Figure 5.2: Loss exhibits a strong log-linear relationship with error detection precision at a fixed threshold, across a broad range of models and hyperparameters (r^2: 0.94; TweetNLP-5, §5.7).

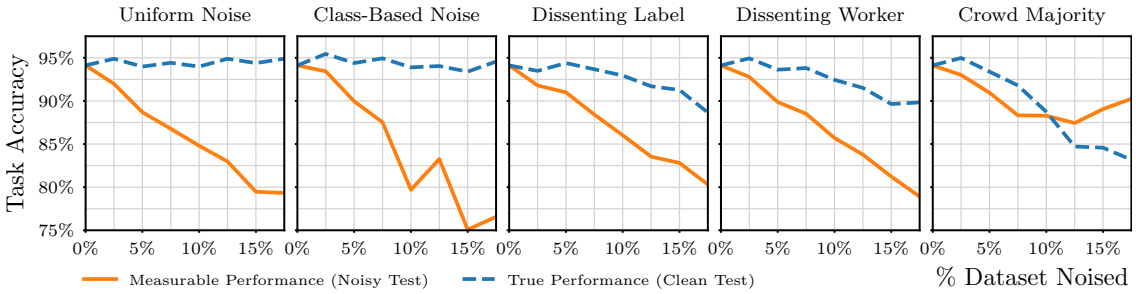

Figure 5.3: Assessing model robustness against a range of noising methods on TweetNLP, with methods ordered by hypothesized realism. Solid orange lines report task performance on noisy test data, reflecting observations in practice; dashed blue lines report task performance on underlying clean test data, reflecting models' actual performance. Models may be robust to uniform and class-dependent noise, where the true performance remains high even with increasing levels of noise. However, they are not necessarily robust to human-originated noise, where the true test performance decreases with increasing noise.

normalized predicted probabilities, also from non pre-trained models, to directly identify label errors. Foundation models are highly performant; we hypothesize that a low likelihood label is likely to be an error.

Foundation models. The success of [Reiss et al., 2020]'s approach in using frozen BERT embeddings motivates directly applying the foundation model paradigm: we use a large language model that was first pre-trained on a task-agnostic dataset, then fine-tune the model for a given task.

We address classification tasks: given a model's score $f_{i,c}$ for each item i and class c, its predicted probability is the softmax-normalized score $p(c \mid x_i)$. Because each item belongs to exactly one class, the contribution of item i to the loss is the negative log probability of the score for the assigned

class y_i:

$$L_i = \sum_i -\log\, p(y_i \mid x_i).$$

We fine-tune such a model for the training split of each data set. To identify label errors on a validation or test set, we hypothesize items from the dataset as a label error in order of the item's loss on that out-of-distribution set.

We propose two main methods. Foundation Model Loss (FML) uses a single foundation model, fine-tuned on the corresponding task (e.g., sentiment classification, POS tagging), to hypothesize items in order of the model-predicted loss. We augment FML using task-adaptive pre-training (TAPT; Gururangan et al., 2020), which is further pre-training on in-domain data, using only text on the pre-training objective without using any labels for fine-tuning on the cross-entropy objective.

Foundation Model Ensembling (FME) combines multiple foundation models on the same task. We hypothesize that ensembling may be disproportionately effective at detecting label errors, as training noise induces models to learn random spurious correlations [Watson et al., 2022]. Rather than using a validation set to choose the single model with the lowest loss on the task, FME uses the top three models trained in a hyperparameter sweep, and differing in both hyperparameters and random initialization, as fully described in Appendix 5.D. FME creates a synthetic probability distribution over the task outputs by averaging the probabilities predicted using each individual model. FME then hypothesizes items in order of loss over the synthetic distribution.

5.4 Generating Realistic Label Noise

To better evaluate label noise detection performance, we prepare a set of benchmark datasets populated with controllable, highly realistic, human-originated label noise.

Sources of human error. We observe that datasets often undergo multiple annotation passes: crowdsourced labels typically aggregate several annotators' inputs [Hovy et al., 2014, Wei et al., 2022], and subsets of data may receive more extensive validation [Bowman et al., 2015], gold labels by trained experts [Plank et al., 2014], or correction passes [Reiss et al., 2020]. We hypothesize that differences between such annotations may be usefully repurposed as a source of realistic, *human-originated* label noise, as disagreements between annotators is known to reflect systematic ambiguity and human error [Plank et al., 2014, Zhang et al., 2017], and differs from the type of noise studied using existing synthetic methods.

We construct three noising methods which may be applied in many of the above scenarios. For any dataset which includes two levels of label quality, the *dissenting label* method replaces final labels with disagreeing labels at random, simulating imperfect quality control. Datasets which provide individual annotator identifiers may apply the *dissenting worker* approach: select one annotator at random, apply all of their labels which disagree with final labels, and repeat until reaching the target

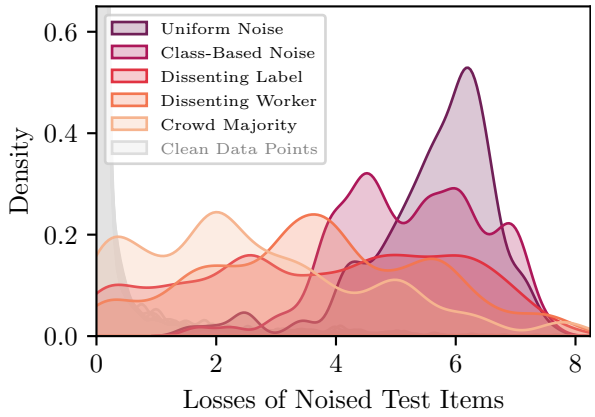

Figure 5.4: Distributions of losses of label errors on TweetNLP at 5% noising. Uniform and class-based noise produce high and distinctive losses; human-originated noise is widely distributed, and has greater overlap with the distribution of clean data points; §6.

noise rate. This simulates gaps in annotator training, which introduce systematic idiosyncrasies. Finally the *crowd majority* method applies to any dataset in which individual annotations can be aggregated to produce a label other than the final label: the former label simulates challenging, systematic errors in the latter.

Noising and robustness. We assess the effect of these noising methods using TweetNLP [Gimpel et al., 2010], a corpus of 26,435 tokens from 1,827 American English tweets collected from Twitter used to train part-of-speech (POS) tagging. TweetNLP includes gold labels annotated by 17 experts, but later received a separate crowdsourced assessment, aggregated by majority vote [Hovy et al., 2014]. We noise TweetNLP to eight levels from 0-20% separately for each method, fine-tune `DeBERTA-v3-base` [He et al., 2021] on each noising, and evaluate models on both noisy and clean test sets. Results from noisy test sets represent model performance as *measurable* in practice; real datasets contend with noise in evaluation data. Clean test set results represent *true* model performance. Fig. 5.3 reports the results of this evaluation.

For uniform and class-dependent noise, true performance remains high even for high noise levels (per [Rolnick et al., 2017]). But crucially, this robustness does not extend to human-originated noise: human label errors are correlated to input text, and so contain systematic erroneous features, which models may learn in training. On more challenging noising methods, although measured performance appears to increase, true performance actually *linearly decreases* with noise. Fig. 5.4 explores this further via the distributions of model losses for each noising method: loss induced by human-originated noise overlaps significantly with clean items, whereas loss from uniform and class-based noising is distinctively higher.

IMDB	New Protocol			
Old Protocol	C	NA	NE	Total
Correctable	105	44	24	173
Non-Agreement	75	252	225	**552**
Non-Error	3	62	520	585
Total	183	**358**	769	1310

Amazon	New Protocol			
Old Protocol	C	NA	NE	Total
Correctable	142	43	117	302
Non-Agreement	140	79	211	**430**
Non-Error	75	31	162	268
Total	357	**153**	490	1000

Table 5.2: Re-evaluation of baselines: The number of **C**orrectable, **N**on-**A**greement, and **N**on-**E**rror assessments produced by the CL Mechanical Turk evaluation protocol and the new protocol, on the same set of items. The new protocol substantially reduces annotator non-agreement; §5.5.

Noise detection benchmarks. We standardize a set of benchmarks from existing datasets for use in our main experiments. TweetNLP-5 and SNLI-5 aim to simulate typical data noise conditions: we apply dissenting worker and dissenting label noising to a 5% level (see Appendix 5.A for details). SNLI is a corpus of 570,152 sentence pairs, in which the task is to label each pair with entailment, contradiction, or semantic independence; we use the 10% subset which includes five crowdsourced annotations per item, as collected by [Bowman et al., 2015] during data validation.

We construct TweetNLP-M to investigate robustness to systematic error introduced by the crowdsourcing process. We apply crowd majority noising, comparing noisy majority-vote aggregated labels by [Hovy et al., 2014] to clean expert labels, which serve as a measure of true performance. Accordingly, we retain all disagreements, or 20.46% of the dataset. We also report results on Recon, a legal classification dataset of 1,279 documents in which [Hong et al., 2021a] found label errors to destabilize model evaluation; as above, we compare non-expert and expert annotator labels.

5.5 Validation on Real Label Errors

In addition to human-originated noise datasets, we evaluate error detection performance on organic errors in two benchmark datasets, following [Northcutt et al., 2021a]'s protocol.

Datasets. The IMDB Large Movie Review Dataset is a collection of movie reviews for binary sentiment classification [Maas et al., 2011], and is split into train and test sets of 25,000 items each. Amazon Reviews is a collection of reviews and 5-point star ratings from Amazon customers [McAuley et al., 2015]. We used the version released by [Northcutt et al., 2021a], which includes the following modifications: It uses 1-star, 3-star, 5-star reviews with net positive helpful upvotes as a ternary sentiment task, resulting in a dataset of 9,996,437 reviews. For tractability we use a train split of a random sample of 2.5 million items, and a test split of 25,000 items.

Baseline protocol. Workers are presented with review text and asked to determine whether overall sentiment is positive, negative, neutral, or off-topic. Each review is independently presented to five workers. An example is considered a "Non-Error" if at least three workers agree the original

	I	Am.	R	T-5	T-M	S-5
H&G	-	-	-	0.30	0.41	0.20
CL	0.24	0.31	0.25	0.30	0.41	0.17
FML	0.58	0.39	0.37	0.66	*0.48*	0.54
FME	**0.60**	**0.40**	**0.38**	*0.68*	*0.48*	0.61
FME+CL	0.20	0.17	0.37	*0.68*	*0.48*	**0.62**

Table 5.3: Main experiment: Evaluating label error detection methods using datasets containing highly-realistic label errors (**I**MDB, **Am**azon Reviews, **R**econ, **T**weetNLP-**5**, **T**weetNLP-**M**, **S**NLI-**5**). Foundation model-based methods significantly outperform baselines on every dataset, as shown by an overall performance metric (AUPR).

label is correct. Otherwise, we consider the label to be correctly identified as an error. We further categorize label errors as "Correctable" if at least three workers agree on the same replacement label, or "Non-Agreement" if no majority exists.

New adaptations. While conducting initial experiments, we found that the [Northcutt et al., 2021a] MTurk protocol resulted in a significant amount of annotator fraud. Some workers spent unreasonably short amounts of time on the text, and frequently disagreed with both expert and peer annotators, reflecting increasingly common issues in crowdsourced annotations [Kennedy et al., 2020]. Appendix 5.C describes four extra conditions we added to improve the [Northcutt et al., 2021a] protocol.

In order to establish an accurate baseline, we re-evaluate the label errors hypothesized by CL [Northcutt et al., 2021a]. On the new protocol, Fleiss' κ inter-annotator agreement increases from 0.131 to 0.464 for IMDB, and 0.014 to 0.556 for Amazon, and Table 5.2 shows that Non-Agreement decreases by 35% in IMDB and 65% in Amazon. This suggests a substantial decrease in low-quality annotations.

5.6 Experiments

Label noise realism. Section 5.4 defined the human-originated noising protocol used to generate TweetNLP-5, TweetNLP-M, and SNLI-5. Section 5.5 specified a protocol for identifying organic label errors present in IMDB and Amazon. We assess the realism of synthetic noise methods by comparing loss distributions against models trained with organic noise (for real label errors, we refer to items verified as Correctable via MTurk). We quantify the degree to which noising induces erroneous learning by measuring the Wasserstein distances between noisy and clean loss distributions.

Overall LLM performance. We assess broad error detection capabilities by evaluating 13 commonly-used LLMs on TweetNLP-5. We measure performance against loss, model size, and GLUE score (a proxy for general model capability; Wang et al., 2018). Appendix 5.D provides implementation

	Precision, Recall @ Error%[3]						Recall @ 2 · Error%			
	I	Am.	R	T-5	T-M	S-5	R	T-5	T-M	S-5
H&G	-	-	-	0.31	0.44	0.22	-	0.54	0.63	0.34
CL	0.41	0.51	0.31	0.36	0.44	0.18	0.46	0.47	0.63	0.32
FML	0.68	0.64	**0.46**	0.65	0.47	0.45	0.62	0.88	0.64	0.66
FME	**0.69**	**0.66**	0.38	0.66	*0.48*	0.46	*0.69*	0.88	0.65	*0.68*
FME+CL	-	-	0.38	**0.69**	*0.48*	**0.47**	*0.69*	**0.89**	**0.66**	*0.68*

Table 5.4: Main experiment: Evaluating label error detection methods using datasets containing highly-realistic label errors (**I**MDB, **Am**azon Reviews, **R**econ, **T**weetNLP-**5**, **T**weetNLP-**M**, **S**NLI-**5**). In practice, estimating the number of dataset errors and checking this many items quickly catches up to 69% of errors, at the same accuracy (Precision, Recall @ Err%). For improved coverage, checking twice this number of items catches up to 89% of errors (Recall @ 2·Err%).

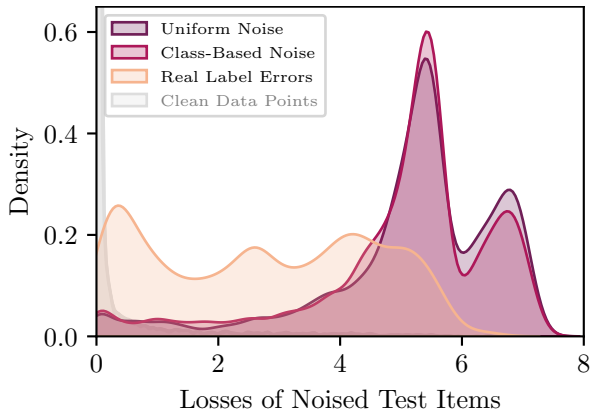

Figure 5.5: Distributions of losses of hypothesized label errors that MTurk workers verified for IMDB. As with Fig. 5.4, uniform and class-based methods do not approximate real, worker-identified errors, and losses of real label errors have greater overlap with the distribution of clean data; §6.

details. This experiment's results inform model selection: we use `DeBERTA-v3-base` for all further experiments.[2]

Main experiment. Using our realistic noising benchmarks, and the MTurk baselines and verification protocol, we can now assess the performance of each label error detection method. We evaluate Foundation Model Loss (**FML**) and Foundation Model Ensembling (**FME**).

As a baseline, we evaluate Confident Learning (**CL**; Northcutt et al., 2021a). CL is not a standalone method; it augments existing models. Given an underlying model's predicted scores for each class and the true proportion of each class, CL forms a reweighting matrix, called the confident joint. To form a label error prediction score, CL reweights the model's scores by the confident joint. CL hypothesizes items in order of this resulting score.

CL uses FastText [Joulin et al., 2017] for IMDB and Amazon, but includes no implementations for POS tagging or NLI. As a result, for TweetNLP and SNLI, we apply CL to the **H&G** baseline [Hendrycks and Gimpel, 2017], a two-layer neural classifier over word vectors pre-trained on a corpus of 56 million tweets [Owoputi et al., 2013]. For all datasets, we also assess applying CL to foundation models (**FME+CL**).

For each dataset, we run 25 hyperparameter sweeps which each fine-tune a model for the given task (e.g., POS tagging) using noisy data, and select the model with the best validation set task performance. We report label error detection performance (not task performance). Area Under the Precision-Recall Curve (AUPR) provides an overall performance score [Saito and Rehmsmeier, 2015, Hendrycks and Gimpel, 2017]. We also report metrics representing performance on competing data cleaning priorities: efficiency requires high precision on a small number of items, whereas coverage requires high recall on a larger number of items. Appendix 5.E.1 describes the Truncated AUPR used for IMDB and Amazon, which are too costly to fully crowd verify.

End-to-end noising. We finally isolate the effects of noise and label error correction for validation and test splits. For each dataset, we prepare three versions of the validation and test splits, respectively: a *clean* version assumed to contain zero errors,[4] a *noisy* version, with label noise deliberately introduced, and a *corrected* version generated from noisy splits using our main error detection method (ranking errors with FME and correcting the top Err% data points). We train 40 hyperparameter sweeps, with performance cross-evaluated on all prepared data splits.

We report three different metrics. We report each model's accuracy on the clean test split as the *true* accuracy. Following the norms of Fig. 5.3, we report the *measurable* accuracy as the accuracy of the model selected using performance on the noisy or corrected validation split on the corresponding

[2]We also use RoBERTa-BigBird for Recon in order to handle its long input passages [Hong et al., 2021a].

[3]Precision and recall are equal when evaluating a number of items equal to the total error count.

[4]For TweetNLP, we justify our assumption in §5.4: expert labels by [Hovy et al., 2014] are considered noise free compared to crowd labels. For IMDB and Amazon, we follow [Northcutt et al., 2021a], which adds several percentage points more noise than naturally occurs.

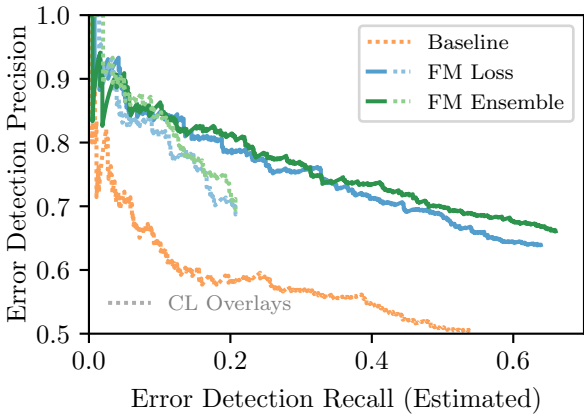

Figure 5.6: Precision-recall curves for label error detection on Amazon by method. FML+CL and FME+CL produce fewer items and do not extend to a recall past 0.21. Applying CL to FM changes little compared to using FM alone.

test split. Finally, we report the *rank* of the model as the rank of the model's performance on clean test data. The best performing model among all sweeps has rank 1, and the worst has rank 40. This metric emphasizes that different validation sets select different models.

We perform this exercise using IMDB and Amazon noised to 5% (I-5, A-5), and TweetNLP-5 and TweetNLP-M.

5.7 Results

Label noise realism. Human-originated noise appears to closely approximate real label noise. Figs. 5.4 and 5.5 show that the losses of both real and human-originated label errors are lower and more widely-distributed than existing noising methods. Their Wasserstein distances to the distribution of clean data are significantly lower than existing noising methods, suggesting comparable erroneous learning (Appendix 5.B).

Overall LLM performance. We discover a strong log-linear relationship between error detection performance and loss, which holds across many model families and configurations (r^2: 0.94, Fig. 5.2). We also find relationships between error detection performance and general model capability, in terms of GLUE score (r^2: 0.79) and model size (Fig. 5.10). Fig. 5.7 illustrates key findings using models' receiver operating characteristic (ROC) curves. Ensembling confers significantly more gains in error detection performance higher than gains on underlying task performance, across a broad range of models and hyperparameters; Appendix 5.E.3 explores ensembling in greater detail.

Main experiment. Table 5.3 shows that Foundation Model Ensembling significantly improves AUPR from the CL and H&G baselines on all datasets, with an absolute difference of 0.36 on

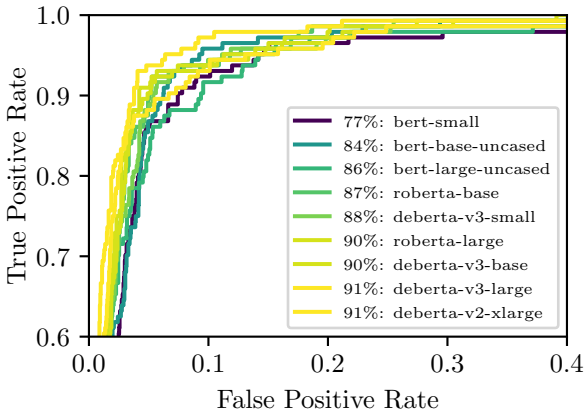

Figure 5.7: ROC curves for error detection performance on TweetNLP-5: LLM loss is highly effective for detecting label errors, and performance is highly correlated with general language understanding (GLUE, r^2: 0.79).

IMDB, 0.09 on Amazon, and a difference of 0.07–0.44 on synthetic data.

Fig. 5.1 shows that applying CL to FME has minimal effect on performance at every level of recall; most numbers are identical across the FME and FME+CL rows of Tables 5.3 and 5.4. In fact, CL does not necessarily improve upon the H&G baseline across datasets, with CL performance sometimes dipping below H&G by 0.01–0.03.

While loss naturally ranks all data points, CL only hypothesizes a fixed number of potential errors: Appendix 5.E.2 shows the raw counts of items at fixed thresholds, per the original CL study. At the CL threshold, we outperform CL by an absolute 15–28%. At the CL+FME threshold, predicted items are almost exactly the same, with Jaccard similarities of 0.59–0.99. By contrast, ensembling improves performance over FML by a greater amount on almost every measure, and introduces no such constraint.

End-to-end noising. Cleaning validation data selects better models. Noise in validation splits reduces performance by encouraging the selection of models with lower true performance. Noise in test splits significantly reduces measurable (noisy test) performance, as expressed by the difference between measurable and true performance. In general, correcting label errors improves task performance: even when the reported task performance worsens, the reported performance is closer to the true performance of the model, measured using clean training and validation data.

5.8 Discussion

Rapid data "health check". Sorting evaluation data by each item's loss is an easy way to quickly highlight label errors. Using this simple technique with a foundation model appears to generally

Eval.	Test Perf.	I-5	A-5	T-5	T-M
Noisy	Measurable	90.1	88.3	89.3	89.3
	True	94.2	91.0	92.8	82.0
	Rank	10	1	3	10
Corr.	Measurable	95.1	90.7	92.9	88.5
	True	95.1	90.8	93.0	82.0
	Rank	4	5	2	8
Clean	True	95.8	91.0	93.8	82.1

Table 5.5: End-to-end effects of label noise on task performance, as evaluated on noisy, corrected, and clean validation and test data splits. True accuracy is measured on clean test sets, and measurable accuracy on noisy or corrected test sets. Rank is a relative measure of true accuracy; lower numerical ranks have higher accuracy. Corrections which improve or reduce performance metrics are highlighted in green or red, respectively. Metrics are evaluated on models trained on noisy data.

identify over half of all label errors through human re-evaluation of a single-digit percentage of all data (Table 5.4). We expect this technique to work across deep learning domains, due to its simplicity and the extensive use of training loss in LNL research [Song et al., 2022]. Given estimates for typical rates of label errors and the gain observed in the end-to-end experiment, our technique may enable a 1–2% increase in reportable test accuracy across many datasets, in addition to the gains from improving model selection.

Pre-training and robustness. We demonstrate that despite established findings on artificial noising [Hendrycks et al., 2018], pre-training confers limited robustness to realistic human noise. The majority of label errors are systematic in nature [Snow et al., 2008, Plank et al., 2014, Samuel et al., 2022], and crowdsourced labels form, to an extent, a different distribution from reality, as approximated by expert labels [Hendrycks et al., 2020]. When trained on crowdsourced or other data containing systematic errors, FMs quickly drift towards this incorrect distribution.

Applying AI to data-centric AI. Data-centric AI aims to improve AI through labeling, curating, and augmenting the underlying data. We find that AI itself can be applied towards improving data quality, as part of a human-in-the-loop (HITL) iteration, which contributes an additional positive feedback loop between data quality and AI performance.

New challenges in LNL. Standard noising methods are unrealistic and no longer challenging for state-of-the-art language models [Algan and Ulusoy, 2020]; recent LNL analyses study conditions where up to 80% of labels are noised [Song et al., 2022]. Our findings reinforce the need to reassess LNL methods in the context of more realistic noise [Zhu et al., 2022].

Our human-originated noising method produces realistic label errors, and can be applied to any crowdsourced dataset which includes raw annotation data. As such datasets emerge across deep

learning domains [Wei et al., 2022], we hope this method may inspire challenging and realistic new LNL performance benchmarks. Our method also enables detailed exploration of the properties of human noise, which may support work on open LNL problems such as improving feature-based noising techniques, and estimating dataset noise [Bäuerle et al., 2022, Northcutt et al., 2021b].

End-to-end noising. The study of model performance on noise in validation and test data is essential: noise in other splits can affect reported model performance as much as noise in training data. Clean and noisy performance on evaluation data provide useful insight into models' overall performance.

Limitations of cleaning benchmark data. In our analysis of model performance gains derived from applying our methods to cleaning evaluation data, we find that cleaning validation splits enables the selection of models with better test performance. Such a method may be useful in a large number of applications.

However, we caution against using this method to clean data intended for use in comparing performance across model families and variants: the cleaning process may bias any such benchmarks toward the models most similar to the model used to clean the data. While our method improves the performance of a given model on a task, and correcting label errors always improves the validity of test data, these improvements is unlikely to improve the performance of all models by the same amount. This limitation is shared with other existing model-based scoring methods such as BERTScore [Zhang* et al., 2020].

5.9 Conclusions and Future Work

Pre-trained models effectively identify label errors on real NLP datasets, definitively outperforming existing methods on the same benchmarks by an absolute 9–36% in AUPR.

Human-originated noising techniques may present a solution to the clear limitations of current LNL noising schemes: they are highly realistic and yet controllable for experimental purposes. We invite further exploration of this family of label noising techniques. We believe human-originated noising enables future advancements across multiple areas of LNL, supporting new tasks and metrics in areas such as the cost of human reannotation, estimation of dataset error, and mitigation of bias. Chapter 4 documents the multiple rounds of annotation that were undertaken to produce labels for the parole hearing dataset, and we advocate for similar efforts to be undertaken for NLP benchmarks more broadly.

Finally, we advocate for LNL to move towards an end-to-end approach of *evaluating with label noise*, which takes into account noise within validation and test splits, and more accurately models the conditions of data in practice. The specific task of information extraction over parole hearing transcripts, as described in the following chapter, serves as a case study of how label errors in a

validation set can impact the model training process. We hope that the analysis serves as an example of how an error analysis can inform not only how to improve the model, but also the underlying data.

Appendices

5.A Noising Benchmarks

This section specifies how noising protocols were applied to create each fixed crowdsourced dataset. Crowd labels for each dataset are available to download from the respective GitHub projects.

5.A.1 TweetNLP-5

TweetNLP-5(%) is a fixed noising of TweetNLP to a 5% noise level in each split. Of the label errors, 80% (i.e. 4% of each split) are assigned using the *dissenting worker* method. The remaining 20% (i.e. 1% of each split) are assigned using the *dissenting label* method. Fig. 5.4 shows that both methods provide similar distributions of label errors. Although the dissenting worker method more realistically captures individual worker idiosyncrasies, the dissenting label method is actually slightly lower loss during training (i.e. harder for a model to distinguish from correct labels).

5.A.2 TweetNLP-M

TweetNLP-M(ajority) directly uses the majority class labels collected by [Hovy et al., 2014] on the Crowdflower platform, which have a 79.54% agreement with the high-quality expert gold labels collected by [Gimpel et al., 2010]. Per the [Hovy et al., 2014] protocol, in the rare case of ties, the tie is broken in favor of the label that matches the gold label, if applicable. Otherwise, a label is selected at random. The "-M" suffix distinguishes the [Hovy et al., 2014] labels from the gold labels.

5.A.3 SNLI-5

The Stanford Natural Language Inference dataset (SNLI) annotations do not include a worker identifier, meaning each item is attached to five crowdsourced labels, but there is no indication of which labels came from the same annotator across the dataset. As a result, we cannot apply the dissenting worker noising method.

SNLI-5 has exactly 5% of its data noised in each split. Of the label errors, 80% (i.e. 4% of each split) are assigned using a method that represents systematic errors, to simulate of dissenting worker method: We use the minority label when there is a 3-2 split between the five labels. The remaining 20% (i.e. 1% of each split) are assigned using the dissenting label method, as in TweetNLP-5.

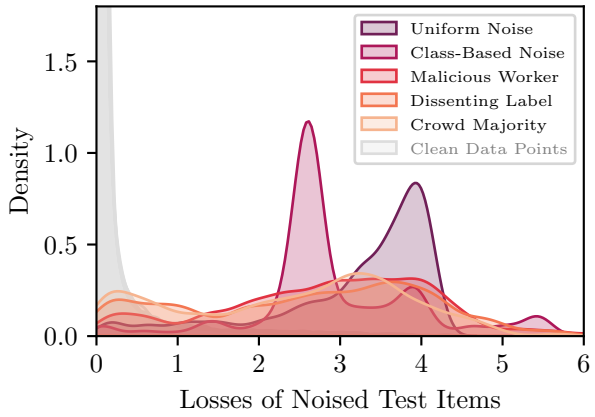

Figure 5.8: Distributions of losses of label errors on SNLI at 5% noising, which demonstrates similar performance characteristics to TweetNLP, as shown in Fig. 5.4.

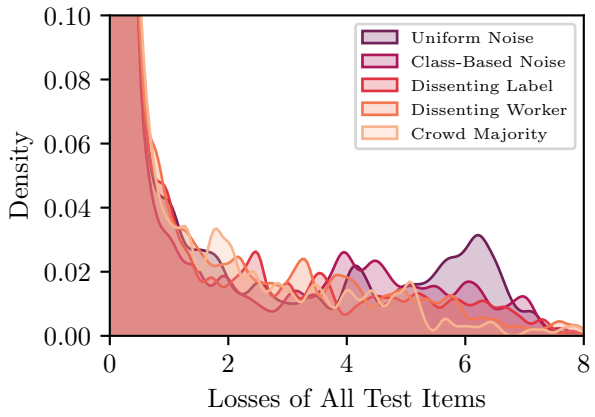

Figure 5.9: Combined distributions of losses of both noisy and clean data points, for TweetNLP with 5% noising.

5.B Loss Distributions

Section 5.4 examines dataset noisings primarily in terms of loss distributions on noised labels. To provide additional context, Fig. 5.8 provides an equivalent view for SNLI, and Fig. 5.9 shows combined distributions of both clean and noisy data points on TweetNLP.

Table 5.6 reports the Wasserstein distances (or earth mover's distances) measured between the loss distributions of noisy and clean data points for models trained on TweetNLP and IMDB, as described in Section 5.7. Human-originated label noise more closely resembles both clean data points and real label noise as its hypothesized realism increases.

Noising Method	TweetNLP	IMDB
Uniform Noise	5.62	5.04
Class-Based Noise	5.23	4.91
Dissenting Label	4.02	-
Dissenting Worker	3.44	-
Crowd Majority	**2.33**	-
Real Label Errors	-	**2.67**

Table 5.6: Wasserstein distances between loss distributions of noisy and clean data points: Human-originated noising exhibits comparable levels of erroneous learning to organic label errors.

	IMDB	Amazon
Original Protocol	0.1314	0.0141
New Protocol	**0.4643**	**0.5561**

Table 5.7: A comparison of inter-annotator agreement between the original and new MTurk protocol results using Fleiss' κ. A score of 1.0 represents perfect agreement between workers, and 0.0 represents guessing at random. Annotations from the original protocol are substantially closer to random chance.

5.C Mechanical Turk Protocol

5.C.1 Change Specifications

We use Amazon Mechanical Turk to validate real label errors from IMDB [Maas et al., 2011] and Amazon Reviews [McAuley et al., 2015]. We begin with the [Northcutt et al., 2021a] protocol, and add four additional conditions, so as to mitigate annotator fraud.

First, we pre-qualify workers by requiring them to correctly answer a qualification test of four unambiguous questions [Hovy et al., 2014, Agley et al., 2021].

Second, after the initial qualification, we continue to monitor worker quality by introducing sentinel questions with known answers into the workers' regular tasks. We periodically remove workers who fail the tasks.

Third, we set filter criteria to limit workers to the following Anglosphere countries: United States, Canada, United Kingdom, Ireland, Australia, and New Zealand [Moss and Litman, 2018], to improve the chances of finding annotators with sufficient cultural context to correctly interpret review text.[5] Our filter criteria include the standard recommendations of requiring a $\geq 99\%$ positive task approval rate with ≥ 500 tasks approved.

Finally, we set a baseline target rate of US$10 per hour, calculated using word counts and average reading speed (primarily for ethical reasons; the effect of compensation and annotation quality is an area of active research; Saravanos et al., 2021).

[5]Despite these precautions, we recognize that every precaution is subject to fraud, e.g., location is subject to VPN and bot attacks. [Dennis et al., 2020, Mellis and Bickel, 2020, Kennedy et al., 2020]

	IMDB	Amazon	Total
Original Correct	33	19	52
New Correct	**41**	**31**	72
Both Correct	28	14	42

Table 5.8: A comparison of original and new MTurk protocol results against 100 expert-labeled data points.

The new protocol's labels are produced using a final set of approximately 70 workers. Workers averaged at least 12 seconds on each task; half the time needed to read prompts at an average reading speed. The average time spent by a worker in the [Northcutt et al., 2021a] protocol was 5 seconds.[6]

5.C.2 Protocol Validation

We hypothesize that the Non-Agreements in the original protocol represent not only ambiguous data points, but also noise in the original protocol resulting from low quality work. Tables 5.2 and 5.7 show that the new protocol improves the level of agreement between workers. As such, we confirm that the increased agreement between workers in the new protocol results from higher quality labels.

Following the [Northcutt et al., 2021a] protocol for expert review, we additionally select a total of 50 items from each of IMDB and Amazon for expert review. The experts are blinded to both the original labels and MTurk results and asked to label each item from scratch. They then reconciled results and came to a consensus for each item. The results are compared at the aggregate level of "Correctable," "Non-Agreement," and "Non-Error," as opposed to the individual sentiment level (Positive, Negative, Neutral, or Off-Topic). The expert agreement with one another was 79%, so in 21% of the items, the expert label was considered to be Non-Agreement and matched the MTurk workers only if the workers also produced Non-Agreement. Table 5.8 provides the result of this assessment.

For the original protocol, 52% of the items agreed with expert annotators, 31% of the items were incorrectly labeled as Non-Agreement, 12% of the items were incorrectly labeled as Correctables, and 5% of the items were incorrectly labeled as Non-Errors. 8% of items were disagreements between experts and crowd workers where neither side had a Non-Agreement. In other words, 8% of all items were disagreements between Correctable and Non-Error.

For the new protocol, 72% of the items agreed with expert annotators, 4% of the items were incorrectly labeled as Non-Agreement, 7% of the items were incorrectly labeled as Correctable, and 17% were incorrectly labeled as Non-Errors. 5% of items were disagreements between experts and crowd workers where neither side had a Non-Agreement.

[6]The reported time is an *upper* bound on the average time a worker spends on a task.

Dataset	Num. Errors Hypothesized	Correctable			Non-Agreement			Non-Error		
		CL	FML	FME	CL	FML	FME	CL	FML	FME
IMDB	1310	183	323	**328**	358	573	581	769	414	401
Amazon	1000	357	508	**517**	148	131	143	495	361	340
TweetNLP-M	250	121	158	**165**	-	-	-	129	92	85

Table 5.9: The number of each type of error accurately identified for each dataset by each noise detection method, keeping the number of errors hypothesized fixed for ease of comparison. (TweetNLP is expert reviewed and by construction does not have any Non-Agreement types.)

Dataset	Num. Errors Hypothesized	Correctable		Non-Agreement		Non-Error		Jacc. Sim.
		FME	FME+CL	FME	FME+CL	FME	FME+CL	
IMDB	316	168	168	108	108	40	40	0.99
Amazon	381	**226**	204	65	56	90	121	0.60
TweetNLP-M	129	93	**98**	-	-	36	31	0.59

Table 5.10: Examining the performance of overlaying Confident Learning on FME, comparing the number of errors hypothesized by FME+CL. We also report the Jaccard similarity between the two models.

5.D Overall LLM Performance Experiments

Due to the high costs associated with expert and crowdsourced validation, we use TweetNLP-5 as a development dataset for model selection.

We selected the following models for exploration: XLNet (`base`, `large`), RoBERTa (`base`, `large`), BERT (`small`, `base`, `large`), DeBERTa (V3: `xsmall`, `small`, `base`, `large`, and V2: `xlarge`, `xxlarge`), GPT (assorted). We performed 25 hyperparameter sweeps with each model, selecting the top three runs for further analysis. In order to avoid model family-level bias in the choice of hyperparameters, we set a broad shared range for three hyperparameters: learning rate varying from 10^{-6} to 10^{-3}, the number of epochs from 2 to 8, and the batch size between 8, 16, 64, and 128. Training time and the final hyperparameters varied based on the model.

We ultimately selected `DeBERTA-v3-base` as a compromise between performance and training speed. We used Google Cloud Platform for training infrastructure. Experiments were run using NVIDIA A100 GPUs, and runtime per training run was approximately 20 minutes for IMDB, Recon, and SNLI, 3 minutes for TweetNLP, and 4 hours for Amazon, when configured with a 2.5 million data point training split.

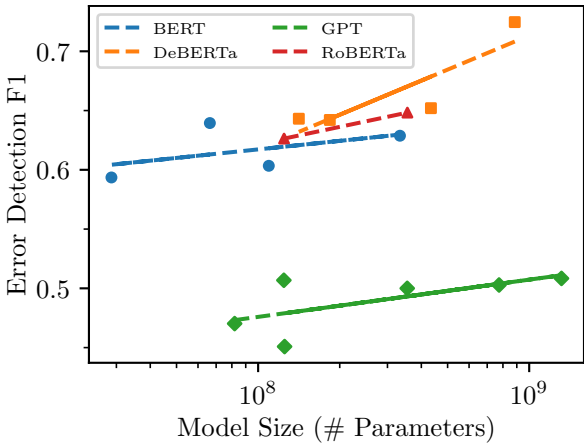

Figure 5.10: Label noise detection performance by model size and family, evaluated on TweetNLP-5. GPT-based models exhibit similar scaling trends, despite intrinsic disadvantages on classification tasks (due to pure autoregressive pre-training).

5.E Main Experiment

5.E.1 Metrics

We calculate the Area Under the Precsion-Recall Curve (AUPR) using the trapezoidal rule, given individual measurements of precision and recall at every possible threshold.

We report the Truncated AUPR on IMDB and Amazon. Because IMDB and Amazon are too expensive to fully crowd verify, we cannot calculate precision and recall at the 25,000th item for each method, for each dataset, as it would require every data point to be relabeled on MTurk. Instead, we use the CL framework of predicting a fixed number of items. For example, for IMDB, CL hypothesizes 1,310 out of the 25,000 items to be label errors. We can calculate the precision and recall for every threshold, up to the number hypothesized by Confident Learning. We can calculate the precision and recall of the 1st, 2nd, 3rd, . . . , and 1,310th items.

We know the exact recall for all synthetic datasets. For IMDB and Amazon, we use the estimate that 5% of the data is erroneous, which is consistent with common understanding of the prevalence of label errors [Redman, 1998, Müller and Markert, 2019, Northcutt et al., 2021b, Kreutzer et al., 2022].

We choose to use the AUPR and truncated AUPR metrics over the more commonly-used AUROC (Area Under the Receiving Operating Curve) because of limitations in measuring recall in practice. Determining the true recall of a label error detection method on a real datasets is generally infeasible due to its high cost; this requires a complete re-evaluation so as to identify every label error within the dataset. While some datasets exist in which this has been undertaken, such as [Hovy et al., 2014] for TweetNLP, for most datasets containing organic label errors, we can only assess precision

Method	Task Accuracy		FM Error Detection Performance		
	Noisy	Clean	Precision	Recall	F1
Averaged	0.88 ± 0.03	0.91 ± 0.03	0.50 ± 0.11	0.65 ± 0.03	0.56 ± 0.08
Ensembled	0.89 ± 0.02	0.92 ± 0.03	0.56 ± 0.12	0.62 ± 0.03	0.58 ± 0.08
Difference	+1.14%	+1.24%	+12.52%	−4.31%	+4.66%

Method	Effects of CL Overlay		
	Precision	Recall	F1
Averaged	0.51 ± 0.11	0.67 ± 0.03	0.57 ± 0.07
Ensembled	0.58 ± 0.11	0.65 ± 0.03	0.61 ± 0.07
Difference	+13.89%	−2.98%	+6.03%

Table 5.11: Ensembling confers gains in error detection performance disproportionate to gains in underlying task performance, across a broad range of models and hyperparameters (on TweetNLP-5, results from top three models per sweep, as measured at the fixed threshold set by CL).

directly.

For IMDB and Amazon, we estimate recall by estimating total dataset error counts using sampling techniques, which are inherently imperfect. Errors in recall scale estimates of AUPR by a fixed ratio, and therefore comparable between models on the same dataset, whereas AUROC is nonlinear with respect to the estimate of recall.

All results reported on synthetic datasets reflect the average of individual scores from the three top-performing models from 25 hyperparameter sweeps. However, for cost-efficiency, results which require crowdsourced evaluation (such as IMDB and Amazon) are based on one run selected at random from a top three.

5.E.2 Confident Learning

Northcutt et al. [2021b] reports results using raw counts, not the accuracy, precision, recall, or any other metric. For ease of comparability, Table 5.9 reports the number of correctable, non-agreement, and non-error items identified by each method on each dataset. CL hypothesizes a fixed number of items, which is reported in the last column, and we assess a matching number of items from each method.

When hypothesizing a fixed number of items, the foundation model approaches far outperform CL baselines. On IMDB, FME correctly identifies 909 label errors, a 28% absolute improvement in accuracy. On Amazon, the FME approach correctly identifies 660 label errors, compared to the 505 identified by CL, a 15.5% absolute improvement.

Applying CL to FME results in a different model that hypothesizes a different number of items (fewer, in all cases). Table 5.10 shows the raw counts of correctable, non-agreement, and non-error items when each of our models hypothesizes items at this reduced threshold.

Overlaying CL on foundation model loss appears to have little marginal utility. Table 5.10 also shows a high Jaccard similarity across all datasets, suggesting that applying CL on top of an FM changes little about the items hypothesized. On many datasets, FME and FME+CL perform almost identically in the number of items correctly hypothesized, slightly harming performance on Amazon Reviews, and slightly improving it on TweetNLP-5 (Table 5.11). FME+CL decreases the total number of hypothesized items compared to FME because of the threshold set by CL. We compare the FME and FME+CL approaches at the reduced number of hypothesized items in order to assess the impact of CL in the presence of pre-training.

Not only is aggregate performance nearly identical, we see in Figs. 5.1 and 5.6 that FME and FME+CL perform similarly for the *entire range* of items hypothesized along the Precision-Recall curve. The primary difference is that FME can continue hypothesizing items even past FME+CL's threshold.

5.E.3 Ensembling

Results from Tables 5.3, 5.4, and 5.11 show that ensembling (FME) improves error detection performance over using a single model (FML) in almost every scenario tested, at a rate several times higher than gains to underlying task performance.

We also observe a phenomenon of disproportionately high variance in model error detection performance: Table 5.11 quantifies the standard deviation of the former at three times the standard deviation of performance on the underlying task, and Fig. 5.2 shows this to be the case even when comparing models with a fixed loss. This finding persisted even when holding all hyperparameters and data constant, with only the random seed being changed.

We hypothesize that label noise in training data induces models to learn spurious correlations, which cause models to make errors in a structured manner [Watson et al., 2022, Jiang et al., 2022]; this results in greater levels of model disagreement, with minimal impact on top-line performance. Ensembling may be disproportionately effective because it serves an added function of reducing variance caused by these low-quality features.

5.E.4 TAPT

We perform Task-Assisted Pretraining (TAPT; Gururangan et al., 2020) using the original hyperparameters everywhere except for the optimizer, in which we use AdamW instead of Adam for DeBERTa. We run TAPT on the all splits of the corresponding data for all datasets except Amazon Reviews, where because of its size, we use TAPT on only 50,000 data points, or 0.5% of the full dataset. After running TAPT, we then run 25 fine-tune sweeps.

Chapter 6

Information Extraction for Parole Hearings

Advancing information extraction and question answering is a major technological step toward the development of the Recon Approach, a paradigm for applying machine learning to criminal law in a way that centers human discretion. To better understand the importance of natural language processing, we use the distinction between *codified justice* and *equitable justice* established in Section 2.2. The criminal justice system struggles to balance "the value of treating like cases alike, and the value of treating each case individually." [Bell et al., 2021] In criminal law, machine learning has been proposed as a tool to improve consistency in decision making, but to date, research efforts have primarily focused on *codified justice* – processes that make a determination given a limited set of case factors and using specifiable rules, such as a risk assessment used for a probation classification. When quantifiable historical data over which the rules have been applied is available, machine learning models can be trained to predict a desired outcome. However, various legal contexts balance a standard of codified justice with a standard of *equitable justice*, which requires decision-makers to apply moral principles to individuals' unique situations.

How can natural language processing aid equitable justice? Equitable justice centers human discretion and the uniqueness of each individual, but nonetheless is based on factual information. The facts of each case are typically discussed and interpreted through dialogue, whether at a full criminal trial, or at a shorter hearing before a judge or other decision-maker. Often, the dialogue produces transcripts, which are available as public records. Usually, the sheer length of transcribed conversational text all but prohibits any meaningful form of quantitative review, because of the immense effort involved in manually annotating case factors. NLP methods for information extraction over speech can assist in identifying the underlying facts of a case from hearing transcripts. The factors can then be used in statistical analyses of a decision-making process to (a) provide historical

understanding over case records that are otherwise locked away in a filing cabinet, and (b) identify specific outlier cases for reconsideration of fair and equitable decision-making where human capacity for review is constrained. By applying information extraction post-hoc rather than filling in a data table or computing a risk score at the time of a hearing, the decision-maker retains full autonomy in conducting a legal process using their own discretion. In this role, information extraction supplements, but never fully supplants, the need for dialogue and transcripts. A broad set of stakeholders can then contribute to identifying the factors that may be relevant in comparing cases.[1] This is as opposed to, for example, using tabular data to compute a risk score that the decision-maker relies on to make a decision.

Existing approaches such as the Predictive Approach described in Section 2.1 generally use tabular data for predictive metrics. An alternative approach is needed for matters of equitable justice, where individuals are judged on a case-by-case basis, in a process involving verbal or written discussion and interpretation of case factors. Such discussions are individualized, but they nonetheless rely on underlying facts. Information extraction can play an important role in surfacing these facts, which are still important to understand. As discussed in Section , information extraction and question answering are important not only for analysis of past cases, but to empower ongoing review of future decisions as integrated into existing oversight bodies.

We present a case study of the capabilities of information extraction methods for dialogue and identify areas for further research in the criminal law context, using the nearly complete dataset of 35,105 parole hearing transcripts for individuals serving life sentences between 2007 and 2019 that we have obtained from the State of California.

Compared to other legal dialogue settings that could be analyzed, the California parole hearing system serves as a useful case study because (1) California has one of the largest prison systems in the U.S., (2) the hearings are transcribed and available on the public record, and (3) the hearings are one continuous dialogue in a single sitting between a decision-maker and a parole candidate, with brief statements from the candidate's attorney. In comparison, criminal trials are much longer, present many forms of exhibits which are often not digitally available, and contain many additional complexities, such as more speakers, cross-examination, evidentiary exhibits, etc.

Our corpus is representative of many challenges in criminal law: (1) Parole hearings, around 20,000 words on average, are longer than documents in existing benchmarks. (2) Existing benchmarks source from written text; parole documents are loosely-structured dialogue. (3) Existing benchmarks contain at least an order of magnitude more labels; the human annotation required for the parole corpus, described in Chapter 4, is expensive. (4) Information extraction from formal written documents centers around named entities and relation extraction. By contrast, much of the text in the criminal context serves the purpose of surfacing, discussing, and correcting case factors,

[1] Bell et al. [2021] describes this approach in the context of the parole system in California. We provide a discussion of the ethical implications of our work in Section **??**.

which are not necessarily relational. This means parole hearings pose both extractive and abstractive tasks, often across multiple sentences, which is known to be challenging even in more structured settings [Wang et al., 2021].

The following sections present two different approaches to tackling challenges (3) and (4)—the challenges of long document length and of scarce labeled examples for training.

Section 6.1 employs data programming Ratner et al. [2016] to solve the two primary challenges. Instead of annotating parole hearings one at a time, SMEs instead write heuristic *functions*, which can be easily computed over all hearings. The data programming approach addresses the scarcity of training data by trading off quality and quantity: every document is assigned at least one label for each feature, but the label generated by the heuristic may be incorrect.

Section 6.2 use a similar approach to the two-step open-domain question answering approach [Chen et al., 2017, Das et al., 2019] by using a Reducer to extract relevant text segments and a Producer to generate both extractive answers and non-extractive classifications. Here, SMEs to write heuristics for only the Reader stage, which we call the Retriever, because of its function in retrieving the most relevant passage from a long parole hearing. Having retrieved a short passage, we can now use relatively sophisticated transformer models for precise question answering over the passage. In a context like ours, with limited labeled data, we show that a superior approach for strong performance within limited development time is to use a combination of a rule-based Reducer and a neural Producer. We study four representative tasks from the parole dataset. On all four, we improve extraction from the previous benchmark of 0.41–0.63 to 0.83–0.89 F1.

6.1 Challenges for Information Extraction from Dialogue in Criminal Law

In this section, we analyze unsupervised, weakly supervised, and pre-trained models' ability to extract such factual information from the free-form dialogue of California parole hearings. With a few exceptions, most F1 scores are below 0.85. We use this opportunity to highlight some opportunities for further research for information extraction and question answering. We encourage new developments in NLP to enable analysis and review of legal cases to be done in a post-hoc, not predictive, manner.

We have identified 11 case factors representative of the types of features (binary, multi-class, date, and numerical) that are relevant to the parole decision-making system and illustrate a range of challenges in information extraction. We evaluate three families of models on this task: (1) an unsupervised data programming paradigm [Ratner et al., 2016] extended to weak supervision, (2) pre-trained question answering models based on DistilBERT Sanh et al. [2019] and Longformer Beltagy et al. [2020], and (3) classification models based on BERT [Devlin et al., 2019] that are each fine-tuned to predict a single task.

Most models fall below an F1 score of 0.85 for most of the features. The different feature types challenge each of the models in different ways. Data programming remains a largely rule-based approach and works best when the keywords indicative of a label are clear, such as the penal code or a numerical education score. Pre-trained question answering models maintain or improve performance on most categories, except for Boolean questions, which remains an area of active development. Surprisingly, all models perform poorly on extracting the risk assessment score, which relies on three simple keywords "low," "moderate," or "high."

Information extraction from long dialogues remains an open challenge, especially when the extraction tasks are not entity-based. We call on research in information extraction to move beyond entity-based tasks in order to tackle the range of tasks relevant for legal dialogue. We also emphasize the need for all methods to handle longer context windows. Long context windows are not merely a byproduct of underdeveloped retrieval methods; they are inherent to the level of personal detail required to apply equitable justice.

6.1.1 Related Work

Information Extraction and Question Answering

Information extraction spans a number of tasks, but neural approaches have concentrated on binary relation extraction. Many relation extraction tasks are performed on only the sentence level [Nguyen and Grishman, 2015, Adel et al., 2016, Levy et al., 2017, Karita et al., 2019, Luo et al., 2019], but techniques have emerged for cross-sentence or even document-level relation extraction [Yao et al., 2019]. Compared to information extraction, question answering allows for a greater range of tasks, represented by the diversity of question formulations [Rajpurkar et al., 2016] and is an alternative approach to the task of creating parole hearing annotations.

For both information extraction and question answering, current top-performing models are pre-trained large language models [Devlin et al., 2019, Radford et al., 2019] that have been fine-tuned on specific tasks, such as question answering.

Applications to dialogue focus on entity-based tasks like argument extraction [Swanson et al., 2015], named entity recognition [Chen and Choi, 2016, Choi and Chen, 2018, Bowden et al., 2018], relation extraction [Yu et al., 2020], and task-based extraction [Fang et al., 2018, Finch et al., 2020, Liang et al., 2020]. Dialogue-like settings are relatively new for question answering. CoQA [Reddy et al., 2019] aims to answer questions over a written text in an abstractive way, but it is only conversational in that multiple questions can be asked of the same source text sequentially. FriendsQA [Yang and Choi, 2019] answers extractive questions about a multiparty dialogue. The questions are considered to be asked of the dialogue, by a third party outside the dialogue. Like FriendsQA, DREAM [Sun et al., 2019b] also uses a dialogue as its source text, but its answers are multiple-choice.

```
Somebody actually took the time to count up all your 115s and make a list
of them for me, and they covered the gambit, but I am very surprised that
you're not a gang member.  We've got attempted murder here in '01, deadly
weapon in '02, battery with a deadly weapon in '05, pruno, '06, mutual
combat, '06, deadly weapon, '06, battery of peace officer, '06.  And
that seems to be sort of the general way your life goes.  You picked up
a couple of these in 2013.
```

Figure 6.1: Example of a section of a hearing during which the deputy commissioner discusses the recent disciplinary history (recorded on Form "115") of the candidate. This occurs about halfway into a 50-page hearing. One extraction task is to identify the date of the most recent disciplinary writeup.

Machine Learning for Criminal Law

Machine learning in law has mainly relied on tabular data, and mostly for prediction, e.g., policing [Ferguson, 2017, Barrett, 2017, Goel et al., 2016], pre-trial detention [Kleinberg et al., 2018a], sentencing [Elek et al., 2015]. Retrospectively, past human (and algorithmic) decisions can be analyzed through the lens of algorithmic fairness, which seeks to understand the way machine learning models or human decisions systematically encode bias [Dwork et al., 2012, Barocas et al., 2017, Corbett-Davies et al., 2017, Corbett-Davies and Goel, 2018, Kleinberg et al., 2018b, Ho and Xiang, 2020].

Within natural language processing, computational linguistics has been used to scale up lexical analyses of various contexts, such as policing [Voigt et al., 2017] and judicial decisions [Danescu-Niculescu-Mizil et al., 2012]. Lexical features can also be used in downstream analysis [Altenburger and Ho, 2019]. Relational information extraction has been applied in the context of using named entities (e.g. attorneys, law firms, judges, districts, and parties of a case) as features for downstream risk analysis for intellectual property litigation [Surdeanu et al., 2011]. However, both extractive and abstractive question answering are still largely unexplored in legal texts.

6.1.2 Data

Our text corpus consists of 35,105 parole hearing transcripts, averaging 18,499 words each, covering 15,852 unique individuals from 2007–2019 parsed from PDF documents. Each hearing is attended by a presiding and a deputy parole commissioner, the parole candidate, and typically an attorney for the candidate. Often, hearings also include a district attorney representative from the county of the commitment offense, who makes a statement, and a victim or their next-of-kin, who may make a statement. Some hearings are attended by visitors who do not participate in the dialogue. The majority of the conversation occurs between the parole candidate, their attorney, and the presiding commissioner.

Feature Selection

We selected 11 features from a set of case factors identified in discussion with legal scholars[2], former parole candidates, advocacy groups including appellate attorneys, representatives from the California Governor's office, and the Parole Board.

Four features are binary: `off mur1` ("Do the controlling offenses include first-degree murder?"), `proggang` ("While in prison, did the parole candidate participate in gang-related programming?"), `da opp` ("Did the district attorney attend the hearing and oppose parole?"), and `job offer` ("Does the parole candidate have an offer letter for a job post-release?").

Two features are multi-class: `edu level` ("What is the parole candidate's education level?"), which falls into one of five categories: "no high school or GED," "high school or GED or CHSPD," "some college courses," "college degree," or "other"; and `risk assess` ("What is the risk score assigned by the psychological evaluation?"), which also has five categories: low, low/moderate, moderate, moderate/high, and high.

Three features are dates. Various dates are mentioned in the course of a parole hearing. Two that are usually stated at the start of the hearing are the MEPD (minimum eligible parole date) and the date that the parole candidate was received into the California Department of Corrections and Rehabilitation (CDCR). Discussing disciplinary writeups that occurred in prison is another key part of the hearing, and we use `last writeup` to denote the year of the most recent such writeup.

Finally, two features are numerical. One is `yrserved`, the number of years the parole candidate has served in state prison. Another is `tabe`, a measure of educational attainment that corresponds roughly to grade levels (10.5 corresponds to finishing half of 10th grade, where 12.9, corresponding to high school completion, is the highest score).

The context window, or section of dialogue required to identify a feature, varies greatly. Figure 6.1 shows an example of a context window for the `last writeup` task. In other hearings, the context window may be longer, e.g., the commissioner may decide to focus on the "mutual combat" in 2006 and speak about the single incident in depth before returning to the list of Forms 115.

Annotation

We collected annotations over a subset of transcripts from three sources. CDCR provided the controlling offense for 26,780 transcripts, which yields `off mur1`. We scraped CDCR's "Inmate Locator" website to obtain `year received` for each parole candidate. Bell [2019] provided human labels for 426 juvenile lifer parole hearings for a superset of the 11 factors.

We manually labeled 827 transcripts with 118 features with a team of 11 research assistants who were trained and supervised by a legal expert. Through the process of annotation, we narrowed down the 118 proposed fields through multiple rounds of annotations and inter-rater reliability

[2]All 11 features are identified as more than marginally predictive in Bell [2019] and Young et al. [2015]'s studies of California parole hearings.

Feature	Num. Train	Num. Val.
off mur1	16,201	1,867
proggang	563	48
da opp	1,173	106
job offer	1,173	106
edu level	1,174	106
risk assess	1,173	106
mepd	1,174	106
last writeup	563	48
year received	10,866	1,261
tabe	367	36
yrserved	982	94

Table 6.1: Training and validation split sizes for each feature.

Feature	Human $\hat{\kappa}$ IRR
off mur1	0.94
proggang	0.93
da opp	0.99
job offer	0.77
edu level	0.92
risk assess	0.80
mepd	0.61
last writeup	0.69

Table 6.2: Inter-rater reliability $\hat{\kappa}$ score of human annotators for each feature

evaluations. The first round of annotations included all 11 features. Subsequent rounds dropped tabe and proggang.

We split data into training and validation sets by sampling at the transcript level. We withheld an additional portion of the data in a separate test split that is not uncovered for the present work in progress. A subset of training transcripts was designated "development" and used for inspection during model development, in particular for developing human intuition for writing label functions.

Because not all features are covered by all label sources, the amount of labeled data varies by feature across the splits. Table 6.1 includes the number of examples in each group.

6.1.3 Human Performance

To compute a human performance baseline for the reliability with which the selected features can be extracted from transcripts, we use Cohen's κ coefficients. Because the overlap of annotators varies by feature, we compute a mean κ-statistic per feature, weighted by the number of documents that overlapped between the annotators . For the kth feature and two labelers i, j, $i \neq j$, let $\kappa_k(i,j) = \frac{p_0 - p_e}{1 - p_e}$, where p_0 is the relative observed agreement among labelers i and j and p_e is the probability of chance agreement under the observed data available for the labelers and let $N_k(i,j)$

be the number of documents for which i and j overlap on feature k. Table 6.2 reports the statistic

$$\hat{\kappa}_k = \frac{\sum_{i \neq j} N_k(i,j) \cdot \kappa_k(i,j)}{\sum_{i \neq j} N_k(i,j)}.$$

6.1.4 Extraction Models

Weakly Supervised Models

Labeling features for parole hearings is burdensome; each hearing takes about one hour to annotate per person. An alternative approach is to generate a noisy but larger dataset using data programming [Ratner et al., 2016]. Data programming improves on purely rule-based methods by learning to automatically weight rules, also known as labeling functions, to produce a probabilistic label. When combined, multiple labeling functions $\lambda_1, \ldots, \lambda_n$ can comprise a high-quality estimate of a single label y. For example, for the task of classifying whether a candidate has a count of first-degree murder, λ_1 can be an indicator of whether the phrase "first degree" appears in the first ten conversational turns. Or, a labeling function might instead rely on neural sentiment analysis models. We wrote a set of labeling functions for each extraction task. We also wrote a retrieval heuristic that selects a number of conversational turns from the transcripts over which labeling functions are run.

We use two strategies to produce an estimate $\hat{y}$ from multiple labeling functions. **Snorkel** MeTaL proposes an unsupervised method [Ratner et al., 2018]. Supervised methods can also be used, e.g. using linear or logistic regression to learn a weighting of the labeling functions to produce an estimate. In our case, we use logistic regression for the binary variables, where learning a prior makes sense, and prior-free constrained least squares regression for all other variables. We call this method weakly supervised labeling functions, or **WSLF**.

Pre-Trained Language Models

Data programming generalizes the knowledge of domain experts; pre-trained language models generalize the knowledge of a large English corpus.

We first use models fine-tuned for question answering, which allows us to use a single model for a wide range of features. We study two question answering models: DistilBERT Sanh et al. [2019] fine-tuned on SQuAD Rajpurkar et al. [2016] and Longformer Beltagy et al. [2020] fine-tuned on SQuAD 2.0 Lee et al. [2020]. We call these two models **QA1** and **QA2**, respectively. Through QA1, we hope to understand the overall performance gain, if any, from pre-training. Through QA2, we hope to understand any advantages of using a model with a longer context window (4,096 tokens) that can handle unanswerable questions, which are common in this corpus.

Our second approach is to model each task as a classification task and to fine-tune a language model for each task. We first fine-tune the base BERT model [Devlin et al., 2019] on all parole hearing text, including unlabeled documents. We then train a classifier layer on the labels produced

	Snorkel	WSLF	QA1	QA2	Task-FT	Avg. # Words
Binary Features						
`off mur1`	0.78	0.74	0.76*	0.78*	**0.80**	974
`proggang`	0.66	**0.87**	0.42*	0.53*	0.64	13,270
`da opp`	**0.83**	**0.83**	0.73*	0.76*	**0.83**	5,219
`job offer`	0.52	**0.63**	0.58*	0.53*	0.46	9,973
Multi-class Features						
`edu level`	0.37	**0.41**	0.13*	0.30*	0.34	12,990
`risk assess`	0.48	0.51	0.46	**0.53**	0.51	12,326
Dates						
`mepd`	0.74	0.83	0.79	0.79	**0.87**	2,405
`last writeup`	0.27	0.03	0.35	**0.42**	0.24	4,811
`year received`	0.47	0.01	0.73	**0.76**	0.15	1,700
Numerical						
`tabe`	0.87	0.88	0.87	0.90	**0.94**	972
`yrserved`	**0.28**	0.08	**0.28**	0.20	0.13	18,603

Table 6.3: F1 scores of information extraction models and the average number of words in the context windows that were the input text for each model. Scores with * in the QA columns required manual intervention to convert the extractive answer into a binary or multi-class label.

in data generation, because of how limited human labels are. We train a separate model for each task (as opposed to a single multi-head multi-task model), i.e. there is one model to predict the binary feature `off mur1`, another one to predict the binary feature `proggang`, and so on. We call this approach task fine-tuned, or **Task-FT**.

6.1.5 Results

Table 6.3 reports the average F1 score across all classes. Binary and multi-class features have natural F1 score interpretations. Date features are quantized into years, and both numerical features have natural quantizations. The TABE score is already quantized to the nearest tenth of a point, and the years served rounded to the nearest year.

Because Snorkel, WSLF, and Task-FT models are trained for a given class, their results are given in the space of the label of the task, whether that is a binary label or a date, for example. However, both QA1 and QA2 models are extractive question answering models, i.e. the answers returned are taken from the text of the hearing. In some cases, the text needs additional processing to be transformed into a label. The transformation may be human intervention, such as in the case of `edu level`, where the extractive answer "ninth grade" and needs to be translated into a categorical answer "no high school or GED." In other cases, such as with dates, the transformation can be partially or fully automated, such as by parsing answers like "March the 6th, 2019" into the MEPD year, 2019, using tools such as SUTime [Chang and Manning, 2012].

Overall, WSLF does well on most classification tasks, though it is beaten by QA2 on `risk assess`

and by the more powerful classifier Task-FT on `off mur1`. QA2 is strongest on dates and generally outperforms QA1. Task-FT performs best on a variety of tasks, but surprisingly, it does not always improve over WSLF and Snorkel, even though its training process uses the very labels produced by the data programming methods, but augmented with even more information, the underlying text itself.

6.1.6 Discussion

Our case study on extracting features from parole hearings illustrates many outstanding challenges in question answering, information extraction, and text classification. Addressing these challenges is key to using NLP for positive impact in criminal law. The tasks posed by the parole dataset do not fall neatly into relation extraction, which has been the focus of neural information extraction. For legal domain tasks, human labels are scarce and expensive, which raises the question of whether weak supervision may be a more efficient allocation of labels than direct supervision. Legal hearings are long and don't fit neatly into the context window size of a neural model, which raises questions about how neural question answering systems can address this task. We answer the questions in turn.

Can weakly supervised methods be successfully used to reduce the cost of data annotation? Data programming provides the opportunity to produce a large number of labels, but it still comes at the cost of requiring experts to translate domain knowledge into programs for each task. Rather than spending one hour labeling one document, an expert may spend dozens of hours designing labeling functions for a single task, e.g. "Does the parole candidate have a job offer?" Once designed, labeling functions are usually computationally light. In producing a final model, adding even weak supervision can improve performance, as seen by improvements of weakly supervised learning functions (WSLF) over the unsupervised Snorkel approach. But unsupervised and weakly supervised techniques mainly perform well only when the tasks can be framed as classification, or when the extractive procedure is relatively simple, such as finding a one-digit decimal TABE score. Reserving some human labels to supervise a WSLF approach outperforms the unsupervised Snorkel method.

Can neural question answering successfully address parole hearings? Neural question answering systems have the flexibility of handling a large range of question formulations and feature types. Compared to other models, this flexibility improves the performance on date features, but surprisingly, on only one additional task, `risk assess`.

Boolean questions remain an outstanding challenge. Reading comprehension datasets, such as CoQA [Reddy et al., 2019] and BoolQ [Clark et al., 2019] include such questions but leave a substantial performance gap for future work. The reliance on manual conversion of some answers to

binary or multi-class labels is problematic.

In general, including on date features, the most common failure mode for QA1 and QA2 is to return an incorrect answer of a correct type. For example, for `yrserved`, the models frequently returned any number they found in the context passage, such as the sentence (e.g. "15 years to life") or any other time range (e.g. "It was around two years I was part of that gang.")

How big a problem is document length? Long context windows continue to challenge all models present, especially neural models. Although developing retrieval models for dialogue can help narrow the context window for downstream question answering applications, an even bigger challenge is the fact that even with an ideal retrieval model, the "correct" context window can still be long. In conversation, speakers are free to go on tangents. More importantly, in the case of legal hearings, speakers elaborate on case factors, attending to detail (as they should), which can greatly prolong a hearing. For example, in discussions of the psychological risk score, both data generation methods and neural question answering systems fail to identify the sentence and keyword containing "low," "moderate," or "high." We suspect that this is because discussions of all risk factors are usually several thousand words long. The score can be mentioned at the very beginning or very end, but often it is tucked away somewhere in the middle.

6.1.7 Conclusion

Parole hearing transcripts go into a great amount of detail in discussing numerous case factors centered around a single named entity, an incarcerated individual who has reached their parole eligibility date. The lack of relational structure and long format of these hearings makes information extraction from transcripts very challenging using several very different approaches from modern NLP.

We estimate that an F1 score of 0.80–0.85 across a broad set of features would provide the ability to conduct meaningful downstream research on a hearing-driven decision-making process like parole. To flag individual cases for reconsideration, we believe that the bar likely lies even higher, since misclassifications often cause outliers. The performance of present models approaches the level at which we can provide useful automatic extraction tools to parole stakeholders for some features, especially certain binary ones. However, for other, seemingly simple medium- and high-cardinality tasks, much work remains.

We plan to conduct future experiments to provide more transparency to model performance. The opaque nature of NLP modeling perplexes our legal collaborators: "How can you identify whether a candidate has participated in gang-related rehabilitation programming but not pick out the risk assessment score from a choice of three words?"

The largest challenge moving forward remains natural language understanding in the face of document length. Of course, length is not the only problem and other artifacts of spoken dialogue

cause challenges, including interruptions, corrections, and colloquial speech. Improved retrieval techniques or even summarization methods can help assess the extent to which document length remains a challenge and possibly mitigate its impact. However, there is no getting around the level of detail that is regarded as due process.

One solution is to incorporate the hierarchical nature of dialogue [Asher and Vieu, 2005]. Within a discussion about risk assessment, a parole commissioner may ask about various sub-factors, such as mental illness, or behavior toward other individuals in prison. We suspect that the word "low," "moderate," or "high" can appear in any of those sub-topics without referring to the risk score. We hope to conduct further research to assess the need for and viability of a hierarchical model. Conversely, an extractive model sometimes picks up on risk-related words in the sub-topics, rather than returning to the higher level question of the risk scores.

Common sense knowledge will also play a role in solving this challenge. In one section of a hearing, the commissioner says, "And, uh, I note that you – you have both a high school diploma and GED, is that correct?" Over the course of the next eight thousand words, the parole candidate describes his life, from playing sports in high school, to having a child, to the chaos of teenage co-parenting, to night school, to getting married, and to moving cities to protect his children. Later on, the commissioner revisits the record and says, "You've taken some college classes," which the candidate himself failed to mention. In addition to understanding the topics and sub-topics in which education occurs, the `edu level` task benefits from real-life knowledge about educational levels. The WSLF model performs well because of tailored labeling functions that encode information about "high school" and "college."

Finally, cross-sentence reference resolution remains important. In Figure 1, the question of the most recent Form 115 can be answered in a short context window. Yet, extracting the answer requires resolving the reference of "these" in " You picked up a couple of these in 2013."

While the amount of attention to personal detail in these hearings presents the biggest challenge to our extraction models, individualized attention is also precisely what defines *equitable justice*. We hope that the NLP community will take up this challenge.

6.2 Learning from Limited Labels for Long Legal Dialogue

In many judicial processes such as legal hearings and criminal trials, decisions are made as a result of lengthy dialogues, in which case factors are discussed in great detail. To study such dialogues, scholars typically invest immense effort to hand label a small number of transcripts with some case factors; the factors are then used in downstream analysis. In most cases, the sheer length of transcribed conversational text all but prohibits any large-scale analysis of the process. Information extraction over dialogues can assist in identifying the underlying factors of a case from transcripts.

The benefits of information extraction are twofold. Automating the extraction of case factors

means that a historical legal analysis can now be comprehensive, containing all available transcripts, rather than being limited to the several dozen or hundred transcripts that a single researcher can label by hand. The second advantage is to open the door to *counterdata* applications in law [D'ignazio and Klein, 2020]. To date, most machine learning applications in the law have been predictive: given case factors up front, make a prediction of an outcome. In domains where case factors cannot or should not be known prior to the hearing, information extraction can produce case factors *after* a hearing, which enables machine learning to play an alternative role to the role of prediction, the role of oversight [Bell et al., 2021]. In our application, information extraction allows the public to audit the parole process, whose case records are otherwise locked away in a filing cabinet.

To be useful for such downstream research, the consensus in legal domain NLP is that information extraction should produce labels that achieve an F1 of at least 0.80 [Hendrycks et al., 2021, Hong et al., 2021b].

The scarcity of labels and specificity of the domain suggest that subject matter experts (SMEs) can be helpful. On the parole corpus, weak supervision-based data programming approaches [Ratner et al., 2016, Zheng et al., 2019] achieve F1 scores of only 0.41–0.63 [Hong et al., 2021b]. We propose an alternative way to involve SMEs, in which we split the problem into two components: a Reducer model which extracts relevant text segments from a hearing, and a Producer model which generates answers from the text segments selected by the Reducer. Our methods effectively achieve extraction at 0.83–0.89 F1.

We show that using an approach with a rule-based Reducer and neural Producer outperforms other commonly-used approaches. Focusing SME effort on developing rules for the Reducer is thus more time-efficient than requiring SMEs to provide additional target labels, whether manually or via data programming. With quality text segments, a neural Producer model can be effectively fine-tuned on just one thousand labels.

6.2.1 Related Work

A review of data programming literature suggests that semi-supervised techniques might be a good fit for our problem space. Several existing pipelines combine a limited amount of training data, rule-based systems and neural models to achieve strong results on benchmark datasets [Maheshwari et al., 2020] and in various medical fields [Ling et al., 2019, Smit et al., 2020, Dai et al., 2021]. By comparison, weak supervision-based data programming methods tend to focus on bootstrapping in the absence of data [Ratner et al., 2017, 2018], which is a nontrivial performance constraint.

Regardless of supervision strength, an architecture based on rule-based systems may be useful for generating "candidates" as input to downstream neural models; [Zhang et al., 2019] explores the time efficiency of manual labeling compared with rule-writing (via regular expressions) for named entity recognition (NER), where results are compared over a bidirectional LSTM-based classifier, finding that in most circumstances, a combination of rule-based and machine-learning classifiers

optimizes human time investment.

We therefore adopt the approach of using a rule-based system for candidate generation. One new challenges with our corpus is that parole hearings generally center around one individual, so the candidates for downstream models are not named entities, but more loosely defined segments of the hearing. Compared to NER, there is less prior work exploring rule-based methods for more general retrieval and segmentation.

Our goal of achieving 0.80 F1 in an abstractive format is currently beyond the capabilities of state-of-the-art (SOTA) neural models on comparable tasks, only one of which is in the legal domain.

On Natural Questions (NQ; [Kwiatkowski et al., 2019]), SOTA models achieve F1 scores of 0.79 and 0.64 on its long and short answer tasks, respectively. However, NQ is purely extractive and averages only 7,300 words per input. On the Doc2EDAG financial statements dataset [Zheng et al., 2019], the Graph-based Interaction model with a Tracker [Xu et al., 2021b] surpasses 0.80 F1 when extracting events from documents averaging 912 tokens in length, but this SOTA result drops to 0.76 F1 in the longest quartile. On Open-Domain Question Answering, the SOTA Dense Passage Retrieval [Karpukhin et al., 2020] has an extractive top-5 accuracy of just 0.66. For downstream applications, a model must have a robust top-1 accuracy.

The closest comparable legal dataset is the Contract Understanding Atticus Dataset (CUAD) [Hendrycks et al., 2021]. Over CUAD, a SOTA model like RoBERTa [Liu et al., 2019] achieves a lower, and extractive, question answering performance of 0.80 recall at 0.31 precision, representing an F1 score of only 0.45, with documents still averaging one-quarter the length of parole transcripts.

6.2.2 Data

We have obtained a corpus of 35,105 parole hearing transcripts, averaging 18,499 words each from 2007–2019.[3] Each hearing is a dialogue, primarily between one or more commissioners and the parole candidate. Most case factors are embellished with history and context, which is important for the procedure of a parole hearing, but challenging for information extraction. [Hong et al., 2021b] identified eleven fields for information extraction. We study the four fields that the previous study failed to extract with near 0.80 F1: `job offer` (whether the parole candidate has a job offer upon release), `edu level` (the candidate's educational level), `risk assess` (a psychological assessment score), and `last writeup` (the date of the candidate's last disciplinary writeup in prison). Figure 6.2 shows examples of how these four features arise in dialogue. On average, each annotator takes forty minutes to label a transcript. Only 3% of the dataset is labeled: `job offer`, `edu level`, and `risk assess` each have 1,173 training examples and 106 validation examples, whereas `last writeup` has 563 and 48, respectively. The corpus also includes 218 transcripts with labeled spans, i.e. the sentences from which the correct label was determined.

[3]Transcripts may be requested from the California Department of Corrections and Rehabilitation under the California Public Records Act.

```
COMM: Dr.  [REDACT], R-E-D-A-C-T, found you to be a moderate risk
and also diagnosed you with anti-social personality disorder.
```

(a) Example of passage discussing `risk assess`, the Comprehensive Risk Assessment score assigned to a parole candidate by a psychologist during an evaluation conducted leading up to the hearing.

```
COMM: When you were going to school, everything was -- how far did
you get in school?
CAND: Junior high.
COMM: Okay.  Junior high, okay.  And have you gotten education in
prison?
```

(b) Example of passage discussing `edu level`, the candidate's level of education. The passage continues for several more conversational turns, in which the commissioner and the candidate discuss various educational programs.

```
COMM: And -- um -- if you are paroled or -- pardon me -- if you
are deported to [REDACT], what is your plan?
CAND: Well -- um -- I had a couple of offers from there -- um --
I would have to -- uh -- check out the -- um -- [REDACT] center
and maybe they could help me -- you know -- train me to get a job
there and get my life together.
```

(c) Example of passage discussing `job offer`, whether the candidate has a job offer upon release.

```
COMM: So when's your last 115?
CAND: Uh, when they had a -- we had a work -- had a work strike
around here.  That was the last 115 I remember.  I forgot what --
what year it was.
COMM: I'm showing one from maybe January of 2010 with a mattress.
CAND: Oh, I didn't realize it was a 115.
```

(d) Example of passage discussing `last writeup`, the date of the candidate's last disciplinary infraction, or Form 115, in prison.

Figure 6.2: Example passages of the four features we study. The speaker `COMM` refers to the presiding commissioner, and the speaker `CAND` refers to the parole candidate.

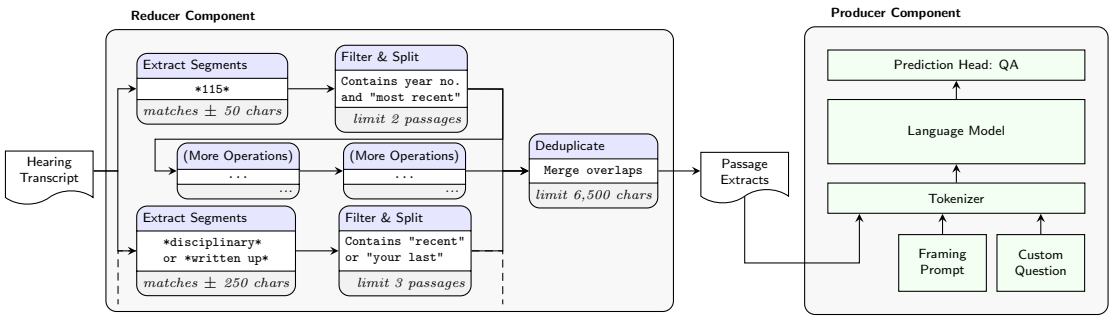

Figure 6.3: Reducer-Producer architecture sketch for the `last_writeup` feature. The Reducer is entirely rule-based, with a few high-level operations over various regular expressions. The Producer is entirely neural and builds on a pretrained language model.

6.2.3 Methods

We use a Reducer-Producer paradigm (Figure 6.3) in the spirit of the Document Retriever-Reader model used in open-domain question answering (ODQA; Chen et al., 2017, Das et al., 2019), with two differences: (1) The Reducer selects one or more relevant passages from within a *single* document [Clark and Gardner, 2018, Krishna et al., 2021], and (2) the Producer model is not necessarily a QA model. We use separate Reducers and Producers for each field. Prior applications of data programming to this corpus used SMEs to write noisy labels for training a neural model; it does not significantly reduce the input text into shorter segments and instead relies on an end-to-end neural approach [Hong et al., 2021b]. By contrast, our approach uses SMEs to focus on the smaller task of reducing input text and relies on only gold labels, however few, for training the neural model. One subproblem is designed to be tractable for an SME (the Reducer), and the other for a pretrained language model (the Producer).

Reducer

The SME (1) encodes keywords and patterns into programmatic rules [Zhang et al., 2019], and (2) evaluates the rules against silver-standard metrics. The SME examines any errors and repeats the process until the development subset is covered to $\geq$95% recall on silver metrics.

Rules. The SME uses keywords to generate candidate segments and candidate substrings (e.g., for risk assessments, "low" is interesting, but only if it occurs in the proximity of "risk"), sequenced in order of increasing breadth and decreasing precision [Zhang et al., 2019]. The framework provides high-level functions that enable SMEs to easily operate on pipelines of candidate segments, filtering in or out, splitting, deoverlapping, and limiting results to create a high-quality reduced output passage.

Evaluation. We reserve the 218 transcripts with labeled spans to serve as a held-out evaluation set. For intermediate SME evaluation and iterations, we use three silver-standard evaluations as a proxy for true Reducer performance: (a) the percentage of results with empty outputs, (b) whether true labels (and common synonyms) appear within reduced passages, and (c) performing interim Producer fine-tuning runs, and evaluating end-to-end performance across a set of hyperparameter sweeps.

Producer

We write several simple rule-based Producers to build an understanding of the problem space, and then fine-tune pretrained language models on the passages returned by the Reducer.

Choice of language model. To ensure high training efficacy, we identify the smallest language model that meets the required benchmark in the general case. We evaluate a range of models' capabilities on a small task: For each of the four fields, we identify ten transcripts with particularly challenging dialogue (see Section 6.C for examples). We manually extract passages from each transcript and benchmark each language model on its average zero-shot classification accuracy on all 40 passages, across 25 random seeds.

Choice of prediction heads. Fields with a small, fixed set of values are a good fit for a classification head (CLS), such as `edu level` which is grouped into four categories, and `risk assess`, for which a psychologist ascribes one of five possible risk levels. Fields with an open-ended set of values may be more suited to the masked language model (MLM) [Hermann et al., 2015, Hill et al., 2016, Chen, 2018, Devlin et al., 2019] or question answering (QA) approach, e.g., `last writeup` can be any year from 1960–2020.

The MLM and QA heads require a user-defined prompt, which are not always natural for all fields. For example, for `job offer`, we prompt MLM with token choices, e.g. "*Commissioner: As to whether you have a job offer lined up: You have [one / none].*"). For `last writeup`, where the correct year exists within each passage, we try various prompts, such as "Your last writeup was in `[MASK]`". We chose prompts with good fine-tuning performance on training data, e.g. for `last writeup`, we use the prompt "Ignoring chronos and 128s, your most recent 115, RVR (rule violation report) occurred in: `[MASK]`"). QA requires a question formulation, for some fields, we augment QA heads with a prefix sentence containing tokens representing all of the current field's possible classes, a technique used in QA benchmarks such as CoQA [Reddy et al., 2019] and BoolQ [Clark et al., 2019], which enables extractive models to always return values from desired classes.

We tried using a multiple choice reading comprehension (MRC) head [Richardson et al., 2013, Lai et al., 2017, Chen, 2018], which proved to be an elegant way of grounding the model, with similarities to contrastive learning, and able to generate dynamic classification options, e.g. unlike

	Prev F1	Train F1	Val F1	Producer Model	Prediction Head
`risk assess`	0.53	0.86	**0.83**	Rules	N/A
`last writeup`	0.42	0.86	**0.84**	RoB+BB	MLM
`edu level`	0.41	0.98	**0.84**	RoB+BB	CLS
`job offer`	0.63	0.96	**0.89**	RoB+BB	QA

Table 6.4: Overall results. Previous best results are from Hong et al. [2021b]. RoB + BB is an abbreviation for the RoBERTa + BigBird model.

year classification, MRC choices are *only* the year that appear in the passage. However, MRC requires a full backpropagation across the entire model for each option of every question, which is memory-intensive for passages where over a dozen options might exist per question, and unnecessarily slow even with gradient accumulation. We do not include MRC in our results.

Training details. We use base models from the HuggingFace Transformers library [Wolf et al., 2020], applying standard hyperparameter ranges [Sun et al., 2019a] and techniques for training BERT-based models, such as the use of a slanted triangular learning rate. However, we set batch size to 1 and use gradient accumulation to simulate a larger batch size, in order to allow Reducer outputs to be as large as possible (approximately 1,500 tokens for RoBERTA + BigBird Base on a 16GB GPU) without affecting training performance. We ran hyperparameter sweeps for approximately six hours per field on a NVIDIA Tesla V100 GPU.

6.2.4 Results

Our methods achieve the 0.80 F1 benchmark[45] for all four fields, as shown in Table 6.4. One rule-based Producer achieved an F1 of 0.83 for `risk assess`, which narrowly outperformed RoBERTa + BigBird model performance of 0.81 F1. However, all other rule-based Producer attempts fell near or below the "Previous F1" mark on their tasks. The `risk assess` task lends itself to rule-writing, because its values are restricted to combinations of "low," "moderate," and "high", and there are a few phrasings that are commonly used (e.g., "Overall, your risk was low to moderate"). By comparison, neural models may have been confused by the multiple other types of psychological assessments that occur in the text (e.g., PCL-R, HCR-20, LS/CMI), which are all assessed on the same "low," "moderate," and "high" scale.

	Rouge-L Recall	Rouge-2 Recall	Bag-of-Words Recall
`risk assess`	0.85	0.76	0.88
`last writeup`	0.87	0.76	0.91
`edu level`	0.92	0.82	0.95
`job offer`	0.87	0.72	0.92

Table 6.5: Evaluating Reducers on labeled spans: Rouge-L and Rouge-2 Recall, Bag-of-Words Recall.

Standalone Reducer Performance

Table 6.5 shows the Reducer's performance on three different measures of recall on the 218 labeled spans. We focus on recall because a Producer can still perform well on a short input even if there are occasional spurious phrases. Also, correct answers are not necessarily unique; labeled spans often point to a single sentence, whereas a fact may be repeated multiple times during the course of a hearing. The Reducer may select *a* correct span, but not the exact sentence selected by the annotator. Recall sidesteps the former issue and slightly mitigates the incorrect penalty imposed by the latter, as similar words may be used in both spans.

The Rouge-L recall ranges from 0.85–0.92: the Reducer frequently finds the exact set of sentences annotated by a human labeler. The Rouge-2 recall is lower, from 0.72–0.82: when the Reducer fails to find the exact sentences, the phrasing of its result is different. However, the bag-of-words recall is still high: 0.88–0.95, which means that the Reducer tends to finds sentences that use almost the same words, if not in the exact same order.

Given the span labeling issue described above, Table 6.5 is almost certainly an underestimate of Reducer performance. This is supported by other assessments of Reducer performance: end-to-end F1 scores of 0.83–0.89 are effectively a guarantee on the lower bound of Reducer performance, and based on the error analysis in Section 6.2.4, only a small fraction of errors were due to the Reducer. This implies *significantly* higher true recall scores. This is also in line with our silver-standard Reducer evaluations, which are consistently above 0.95.

Language Model Benchmarks

Benchmark performance for each language model is provided in Table 6.6. Figure **??** plots model performance against size and shows power-law scaling characteristics, a known feature of neural language models [Kaplan et al., 2020].

Given the relatively small range in performance between the models in our evaluation set (7.5% across all model families and variants), we also run some supplementary tests, finding that (a)

[4]F1 scores are calculated on *exact match* for all prediction heads, instead of the relatively easier bag-of-words metric used in the extractive setting, or precision at 0.80 recall [Hendrycks et al., 2021]. This is a more accurate measurement of abstractive performance, which is essential to downstream results.

[5]Related existing work reports F1, but F1 is an imperfect proxy for the impact of errors for downstream analyses. Any application that seeks to use extracted data should perform its own analysis to understand the relative costs of, for example, false positives versus false negatives for a given field.

Model Family	Model Variant	Size	Max Length	Benchmark Score
BERT	Base Cased	108M	512	26.7 ± 9.4
	Large Cased	334M	512	33.0 ± 8.0
RoBERTa	Vanilla Base	125M	512	29.8 ± 8.1
	Vanilla Large	355M	512	28.3 ± 8.7
	BigBird Base	**128M**	**4,096**	$\mathbf{30.5 \pm 5.6}$
	BigBird Large	360M	4,096	33.0 ± 6.0
Transformer-XL	Vanilla Base	284M	N/A	32.1 ± 7.6
	XLNet Base	117M	N/A	29.1 ± 5.9
	XLNet Large	361M	N/A	30.2 ± 5.9
GPT	GPT2 Base	124M	2,048	32.2 ± 6.3
	GPT2 Medium	355M	2,048	33.5 ± 5.6
	GPT2 Large	774M	2,048	34.0 ± 7.1
	GPT-Neo 1.3B	1.32B	2,048	34.2 ± 6.3

Table 6.6: Zero-shot language model performance (average classification accuracy) on a benchmark of complex, challenging passages, over 25 random seeds.

Hyperparameter	Value(s)
Learning Rate	5e-4 to 5e-7
Batch Size (Accumulative)	1, 2, 4, 8, 16
Number of Epochs	6 to 10
LR Warmup Epochs	0.4, 0.6, 0.8, 1.0, 1.2
Dropout	0.1
Adam Optimizer	$\beta_1 = 0.9, \beta_2 = 0.999$

Table 6.7: Hyperparameter sweep configurations for prediction head selection exercise.

models pretrained on question answering datasets performed 10–15% better in this setting, but a comprehensive evaluation was not feasible as QA outputs are extractive and require manual assessment, and (b) large GPT models performed dramatically better in the few-shot setting, with GPT3 performing at 90–100% accuracy on some problems.

We ultimately use RoBERTa + BigBird Base (RoB + BB; [Zaheer et al., 2020]) as our default model due to its balance of long input length, low computation requirements, and performance. This model supports inputs of up to 4,096 tokens, allowing the Reducer to provide multiple candidate passages without having to split input into multiple model calls and integrate a la [Clark and Gardner, 2018]. It is in the smallest size class of the models tested, facilitating the fine-tuning of large input passages within GPU memory limits. Within its size class, RoB + BB is the second-best performer, performing within 2–3% of models 2–3x its size. Compared to the top performer (GPT2), BERT is known to have better versatility on downstream tasks [Klein and Nabi, 2019] and well-explored fine-tuning characteristics [Sun et al., 2019a].

	CLS	MLM	QA
last_writeup	0.76	*0.79*	*0.82*
edu_level	**0.82**	0.43	0.70
job_offer	0.83	0.69	**0.89**

Table 6.8: The effects of different prediction heads on Validation F1 scores (results in italics are not definitive, as MLM outperforms QA on end-to-end evaluation).

Prediction Heads

We evaluate the gains from prediction head choice by performing 25 fine-tuning runs for each combination of field and head and reporting the highest validation F1 score achieved for each. To ensure test fairness within a reasonable amount of computation, each run uses a random configuration from Table 6.7. F1 scores are recorded at the point where validation loss is at a minimum.

Table 6.8 shows the performance of each prediction head on each field. `edu level` and `job offer` performed comparably to the main runs in Table 6.4. `last writeup` performed best under a question answering head during this exercise, but underperformed the masked language model F1 score of 0.84 in Table 6.4, leaving this result ambiguous. Selecting a suitable prediction head dramatically affects model performance after fine-tuning: suboptimal head choices result in F1 scores of 52-93% of the scores achieved with the best prediction head.

The CLS prediction head performs well across all fields except `last writeup`, where only 20% of all runs score above 0.25, and most score below 0.10. Classification is not a natural format for this field: in order to classify a passage, the model must learn 50 separate classes, one for each possible year from 1969–2019. CLS performs well when the number of classes is relatively low, especially when the answer is abstractive. However, it tends to fail to understand factual relationships. For example, when used for risk assessment its ratings correlate with the number of times the word "gang" or "murder" occurs in the passage (see Section 6.D for more information).

The MLM head has nearly the opposite performance characteristics: it performs best on `last writeup`, at an average level on `job offer`, and very poorly on `edu level`. It is telling that `last writeup` can be expressed as a sentence with a single masked token (which may hold many values), whereas the classes of the latter are all concepts which do not fit into a single token. The MLM head's F1 scores tend to be several points lower than its accuracy, a symptom of the model occasionally filling the mask with arbitrary freeform values.

The QA heads perform well on `job offer`, fairly well on `last writeup`, and at an average level on `edu level`. The first field is easily expressed as in the form of a yes/no question, and the second field's value is extractable from within the passage as with a regular QA task. However, the third requires the model to parse the passage to locate the answer, classify this into one of four fixed phrasings, and return this phrasing from the prefix sentence, a task which is somewhat foreign to a question answering-based model.

Error Analysis

Errors fall into a few clear classes. Approximately 70% of all errors result from what appears to be the model learning spurious associations with co-occurring words. For example, in one conversational turn, a parole candidate describes both his own and the victim's level of education. The Producer incorrectly returns the victim's level of education, which uses the phrase "college courses."

Around 10% of errors result from complex passages (comparable to the examples in Section 6.C), which continue to challenge language models. Spoken narrative language can be arbitrarily complex, and grounding in real world knowledge and presuppositions remain hard to encode. In one transcript, the commissioner asks, "Are you working towards a college degree?" which presupposes that the parole candidate completed high school. However, the model classifies this candidate as not having completed high school or a GED, as the transcript does not explicitly mention either. Some passages require numerical abilities which smaller language models tend to find difficult [Dua et al., 2019]. Table 6.6 suggests that a larger language model may improve performance in many cases.

In the remaining 20% of errors, the Reducer failed to find a match for a given transcript or returned an incorrect passage.

Surprisingly, we found that in 15–50% of the total errors returned (varying by field), the model was actually correct, and had identified incorrectly-labeled or ambiguous data. To be conservative, we did not adjust F1 scores upwards and instead excluded the examples from this error analysis.

A detailed breakdown of errors for `edu level` is provided in Section 6.D for illustrative purposes.

6.2.5 Discussion

Combining Rules and Neural Models

Previous approaches to our problem use rule-generated labels to supervise a model. We instead split the problem into two, where the Reducer is entirely rule-based, and the Producer trains only on the few, but high quality, human labels.

Both rule-generated labels and a rule-based Reducer scale with the number of features to extract, but not the complexity of model or dataset. However, given a fixed development time, we find it more valuable for an SME to focus on only the Reducer. In contrast, end-to-end data programming requires rules for the Producer as well, which can be much more challenging to write. On our data, it takes about ten hours for an SME to write Reducer rules for a model that performs at the exceptional recall rates from Table 6.5. [Hong et al., 2021b] report the same number of hours per feature for an end-to-end data programming model, which performs much worse overall.

As future work, we hope to investigate whether a well-designed Reducer can improve human performance in creating gold-standard labels, saving time by reducing the need to read through entire transcripts.

Assessment of Human-in-the-Loop

We find that an hand-written rules can effectively isolate key segments of text in the overwhelming majority of situations.

The tradeoff of incurring the cost of writing rules per each additional feature proved to be very reasonable for our domain. We have few features, and our requirements demand accuracy over speed. In comparison, prior work suggests that for a neural model to achieve accuracy in the same ballpark, the model would require an order of magnitude more spans, which would be a prohibitive cost. In the general case, when applying our architecture, the per-feature cost of SME time should be considered against (a) the potential per-example savings from reducing labeling requirements, and (b) the performance requirements of the problem space.

The Human-in-the-Loop (HITL) approach enables SMEs to exert a positive influence on the quality of both the final model and the dataset. Given a probable baseline label error rate of a few percentage points [Alt et al., 2020, Reiss et al., 2020, Northcutt et al., 2021b], as the Reducer's recall increases towards the 0.9 level, many of the mismatches against silver-standard Reducer evaluations and fine-tuning errors will actually be labeling errors. For example, in a case study where we checked `last writeup` Reducer outputs against a silver-standard evaluation, we found that over 80% of "errors" were actually errors in human labeling. This also provides opportunities for SMEs to apply domain knowledge to more subtle classes of data issues, such as where Reducer rules surface mislabelings caused by labeler confusion.

As such, a unique advantage to HITL over a neural-only model is improving data quality during the training process. Purely neural models are forced to learn from mislabeled data points, which destabilizes benchmarks and damages model performance. [Northcutt et al., 2021b] By comparison, we frequently detect label errors prior to fine-tuning, and as errors tend to occur in patches (such as under a particular labeler or a particular time period) we can quickly make corrections or exclude large bad patches from the training dataset. This can significantly increase training performance: excluding a patch of bad labels resulted in a 0.2 F1 improvement in one case. Section 6.B elaborates on the data quality improvement process.

Modular Architecture

The Reducer-Producer architecture is useful for enabling iterative, componentwise development. Components may be improved in isolation as requirements arise, such as improving Reducer coverage or upgrading Producer language models, heads or prompts, and sometimes may be entirely replaced without any impact to their counterparts.

In particular, we hope to leave the door open for a general neural Reducer and Producer, allowing downstream users to perform open-ended querying and exploration of the dataset. This architecture enables future work to continue to use our Producer models, which are already trained. The information bottleneck between its components allows for rigorous measurement of the quality of

Reducer output, which enables each component to be trained separately. Additionally, using present models to generate silver-standard data labels may alleviate issues of label scarcity.

6.2.6 Conclusion

Our corpus of parole hearings poses the challenge of information extraction with few gold labels: one thousand labels is not enough to locate and identify the answer in a long document. Parole, like many other applications, requires domain-specific knowledge, which raises the question of how best to incorporate the labor of subject matter experts to assist neural models in making optimal use of available labels, in order to achieve high performance on extraction tasks.

We identified two problems with existing work on the parole dataset, which fell short of the 0.80 F1 on many tasks: (1) Text segments remained too long for many SOTA neural models to digest, and contained many spurious signals. (2) Question answering was a useful first approach to handle a wide range of different feature types. However, out-of-the-box, it was rarely the best way to handle each individual feature type.

We present an approach that uses an SME-designed rule-based Reducer to identify relevant text segments, and a neural Producer to generate labels using those text segments.

We argue that it is time-efficient and performant for human SMEs to write mostly keyword-based rules for finding relevant parts of a parole transcript. In a parole transcript, a field of interest might be discussed in practically infinite different ways, but is usually somewhat well-defined by a limited set of words and patterns that are almost always used (for example, “GED”, “college courses”, “did not graduate” for a parole candidate’s level of education). These keywords are relatively easy for a human to identify and write combinations of regular expressions to identify. However, training a neural model to recognize the phrases over the course of 20,000-word documents requires at least an order of magnitude more labels than are available [Hendrycks et al., 2021]. Therefore, we focus SME energy on the Reducer, and *only* the Reducer.

For the Producer model, the role of human and machine are reversed. When the text is shortened to a sufficiently succinct context, neural models can be successfully fine-tuned to extract labels at an F1 of 0.80. It is practically impossible for a human to write rules to interpret every possible phrasing of, for example, someone’s educational journey. However, pretrained language models excel at producing labels from small, targeted pieces of text. The 1,000 available labels are sufficient for good performance on this task [Zhang et al., 2020]. We use a base model that can handle relatively long tokens. We also explore a range of different fine-tuning heads.

Our architecture shows the effectiveness of a modular, two-step approach, where not every module needs to be a neural or machine learning model. Such efforts to involve subject matter experts are especially important in applications that require substantial domain expertise. We hope that this work encourages additional research to better understand other legal processes whose workings are yet opaque to the public.

Appendices

6.A Reducer Operations and Rules

SMEs write Reducers for each field by composing pipelines of high-level operations, as described in Table 6.9. Operations run on an input transcript or a list of text segments, and emit matches which are compiled into a final output passage. Table 6.9 defines the operations at the general level, and to give an concrete example of how these operations are used, Table 6.10 lists the operations for the `job offer` feature.

Extract Segments	
Extracts a list of segments from a raw transcript which match one or more regular expressions (regexes).	
Input	Transcript text with any preprocessing.
Regex	Accepts a list of regexes and searches the transcript separately for each item, returning matches in the same left-to-right order they are found.
Limit	Length of segment returned around each match.
Filter & Split	
Filters a list of segments against two lists of regexes, to return two lists of matching and non-matching segments.	
Regex	Accepts a "filter in" regex list which segments must match, and a "filter out" regex list which segments must not match.
Emit Matches	
Saves segments from a given list to a specified list for future compilation.	
Limit	Length of segment to store around each match, and maximum segments to store.
Deduplicate	
Ensures a list of segments is free of duplicate or overlapping text ranges. Merges segments with partial overlaps.	
Compile Passage	
Merges a list of segments into a single text passage.	
Separator	String inserted between each segment.

(Continued overleaf)

Limit	Trims passage to a maximum length.

Table 6.9: Overview of Reducer operations.

#	Param.	Value(s)
01	**Extract Segments**	
	Input	Transcript (lowercase)
	Regex	`job offer`
	Limit	1,000 chars centered on each match
02	**Filter & Split**	
	Input	Operation 01
	Regex	`letter`
03	**Emit Matches**	
	Input	Operation 02: Matches only
	Limit	2 segments
	Effect	*Emits 2x1,000-char segments which mention "job offer" in proximity to "letter".*
04	**Emit Matches**	
	Input	Operation 02: Non-matches only
	Limit	2 segments, 500 chars centered on each match
	Effect	*Emits 2x500-char segments which mention "job offer" but not "letter".*
05	**Extract Segments**	
	Input	Transcript (lowercase)
	Regexes	`jobs? ([\w,]+ ){2,10}offer OR offer\w+ ([\w,]+ ){2,10}job`
	Limit	500 chars centered on each match
06	**Emit Matches**	
	Input	Operation 05
	Limit	2 segments
	Effect	*Emits 2x500-char segments in which "job" and "offer" are within ten words of each other.*
07	**Extract Segments**	
	Input	Transcript (lowercase)
	Regex	`(?:find\w+|locat\w+|get\w+) (\w+ ){0,5}(?:work |employment|job(?! offer))`

(Continued overleaf)

#	Param.	Value(s)
	Limit	1,000 chars centered on each match
08	**Emit Matches**	
	Input	Operation 07
	Limit	2 segments
	Effect	*Emits 2x1,000-char segments in which a verb and a noun about job hunting are within five words of each other.*
09	**Extract Segments**	
	Input	Transcript (lowercase)
	Regex	`(?:job(?! offer)|employ|hire|work)`
	Limit	1,000 chars centered on each match
10	**Filter & Split**	
	Input	Operation 09
	Regexes	`letter AND` `offer`
11	**Emit Matches**	
	Input	Operation 10: Matches only
	Limit	2 segments
	Effect	*Emits 2x1,000-char segments which contain a word about employment in proximity to both "offer" or "letter".*
12	**Filter & Split**	
	Input	Operation 10: Non-matches only
	Regex	`letter`
13	**Emit Matches**	
	Input	Operation 12: Matches only
	Limit	2 segments
	Effect	*Emits 2x1,000-char segments which contain a word about employment in proximity to only "letter".*
14	**Emit Matches**	
	Input	Operation 12: Non-matches only
	Limit	5 segments
	Effect	*Catch-all: Emits 5x1,000-char segments which contain a word about employment.*

(Continued overleaf)

#	Param.	Value(s)
15	**Deduplicate**	
	Input	All emitted segments
16	**Compile Passage**	
	Input	Operation 15
	Separator	`[SEP]`
	Limit	First 6,500 characters

Table 6.10: Reducer pipeline for `job offer`.

6.B Improving Data Quality using Silver-Standard Evaluations

Mismatches on silver-standard Reducer evaluations were often a product of real label errors: the datasets examined in [Northcutt et al., 2021b] had a 3.4% error rate on average, which is a similar order of magnitude to label errors encountered in our dataset when performing detailed manual verification.

The parole dataset includes records that span over more than a decade, and labeling has occurred in several waves over the years. As such, the semantic meanings of labels includes subtle shifts and inconsistencies. For example, a blank label might mean any one of the following:

- the annotator was uncertain,
- the transcript is unclear,
- the transcript is clear but the situation itself is ambiguous,
- "none" is a reasonable answer in this situation (such as `last writeup` for a candidate with zero writeups),
- the feature was not applicable in this situation (such as `job offer` for a candidate who is not working age); or
- the feature was simply not fully annotated.

To address these issues, we: (a) write code to correct issues where this is possible, (b) drop entire sections of low-quality train labels where patterns of errors exist, (c) hand-correct validation labels and keep track of all manual corrections, and (d) write small data transforms to simplify the job of the Producer (e.g., fixing common spelling and transcription errors).

6.C Sample Challenging Passages

Table 6.11 provides examples of the complex, challenging passages selected to benchmark language models in section 6.2.4, trimmed for brevity and redacted as per the conventions described within

Figure 6.2.

6.D Supplemental Error Analysis: `edu level`

This section provides a detailed breakdown of the error analysis for a single field and data split (`edu level`, Validation), in order to illustrate typical patterns of errors encountered in our fine-tuned models.

This field was fine-tuned with a classification (CLS) prediction head, and correctly classified 89/106 of its labeled examples. Its 17 incorrectly-classified examples are examined in Table 6.12. The four possible values this field may hold are:

- NA: Did not finish high school
- HS: Completed high school or GED
- SC: Some college classes
- GC: Graduated from college

Field	Passages
risk assess	COMM: With respect to violence risk assessment conclusions [...] the doctor uses a number of measurements. One is the PCL, which is the psychopathy checklist, and states that, "Overall score placed Mr. [REDACT] in the moderate range of psychopathy. [...]" Historically, on the HCR checklist, HCR20, the doctor writes, "[...] he has risk factors that place him in the low moderate risk range for future violence [...] The inmate's overall LS/CMI score indicates that he is in the medium category." And then the doctor goes on to discuss the historical domain and concludes, "[...] the inmate presents a moderate risk for future violence. [...] In the clinical or more current and dynamic domain of risk assessment [...] the inmate presents a moderate risk of future violence. As for the management of future risk domain [...] the inmate presents as a low risk of future violence. Overall then, risk assessment estimates suggests that the inmate poses a low moderate likelihood to become involved in a violent offense if released to the free community."
edu level	COMM: Okay. So, and at the last hearing, it was discussed and I don't want to get -- Well, that's parole plans. We're not going to talk about that right now. But, so you've taken a number of courses. It looks like in 2013, 2014, General Studies. Are you working towards a college degree? CAND: No. We're not able to take a college degree where I'm at. COMM: You say you've taken World War II, Europe Civilization, Ecology. Are these television courses or -- CAND: They're videotapes, CDs.
job offer	COMM: Do you have any job offers if you were to get a parole date? CAND: Uh, I used to be a mechanic before in, uh, [REDACT], my not in a company, but uh, in uh, a little shop with my friends. COMM: Do you have any job offers as a plumber? CAND: Yes. No, no, no, no, no, no. Not as a plumber. But, uh, I got, uh, as a mechanic I got offer with my cousin. COMM: Okay. Yeah. But he's in the United States, right? CAND: No, he's in [REDACT].
last writeup	COMM: You've had 19 115s, starting in 1996, and most of these have been covered in prior hearings but, sort of running through them, couple in 1996, two in 1997, four in 1998, two in 1999, three in 2001, 2002, 2004, 2005, there was a pair. And then 2008, disobeying a direct order was your final 115. What was the 2005, knowingly providing a false claim?

Table 6.11: Examples of complex, challenging passages from parole hearings.

#	Source	Type	Label	Pred.	Error Details
1	Dataset	Ambiguous situation	HS	NA	Self-reported overseas high school completion (no records)
2	Dataset	Ambiguous situation	SC	HS	Vocational courses but taken at a college
3	Dataset	Mislabeling	HS	SC	Transcript explicitly discusses college courses taken
4	Reducer	Reducer pattern miss	NA	HS	Did not capture key sentence: *"I loved school. You know, I played the cello, you know, was ahead in school. I was graduating. I needed one credit to graduate from high school."*
5	Producer	Spurious associations	NA	HS	Two discussions about GED
6	Producer	Spurious associations	SC	HS	Confirms receipt of GED twice plus vocational training, just one brief mention of a college course
7	Producer	Spurious associations	NA	HS	Cluster of words: "school", "high school", "GED"
8	Producer	Spurious associations	SC	HS	Three mentions of graduating high school, one brief mention of college courses
9	Producer	Spurious associations	SC	GC	Candidate discusses the future receipt of an Associate's degree, later uses word "degree"
10	Producer	Spurious associations	NA	HS	Mentions "school" twice, "grade" three times
11	Producer	Spurious associations	SC	HS	Confirms receipt of GED twice, vocational training, two mentions of college courses
12	Producer	Spurious associations	SC	HS	Confirms receipt of GED twice, vocational training, mentions two colleges but not classes, units or degrees
13	Producer	Spurious associations	NA	HS	Mentions "school" three times, "college" twice
14	Producer	Spurious associations	GC	SC	Candidate confirms he has been doing college courses and is close to qualifying for an AA degree, but later notes he already has one degree
15	Producer	Spurious associations	SC	HS	Cluster of discussion around hgh school diploma, GED (four mentions) and reading scores
16	Producer	Spurious associations	SC	HS	Four separate mentions of having receiving GED, one small mention of college courses
17	Producer	Complex phrasing	NA	HS	Description is challenging to interpret: *"I started ditching school and hanging out when I was in high school. I think part of the reason for that was because we never had anything at home, everything was always, seemed like we're always struggling for everything, you know. Our electric bill, I didn't want to keep living like that, so I left, I left when I was 13 years old."*

Table 6.12: Example-level error assessments: `edu level`.

Chapter 7

Factor-based Findings about California Parole Hearings

This chapter presents descriptive findings about factors that statistically predict[1] parole hearing outcome. Leveraging the raw and manually annotated data described in Chapter 4 and the NLP-extracted data described in Chapter 6, we are now able to present the most comprehensive description of the California parole hearing system to date. We find that outcomes are disproportionately predicted by multiple factors outside of the candidate's control, such as which commissioner presides over the hearing, whether the candidate is represented by private counsel, and whether the district attorney appears at the hearing. Notably, these relationships are highly significant even when controlling for case factors.

7.1 Background and Related Work

Computational methods have been broadly applied to study various parts of the American criminal justice system, which range from policing [Gelman et al., 2007, Pierson et al., 2020], to pre-trial detention [Arnold et al., 2018], and finally, to sentencing [Klein et al., 1990, Anderson et al., 1999, Abrams et al., 2012]. However, there have so far been no large-scale studies on the decision-making that decides parole. In other words, much of the existing empirical literature has focused on the many decision steps involved in how the criminal justice system incarcerates individuals. However, comparatively little academic, and even political, attention has focused on the processes that determine how individuals are released from prison.[Bynum and Paternoster, 2019] One national large-scale study of parole uses National Corrections Reporting Program data to understand broad legal and political trends among 237,781 individuals, but only includes those who were conditionally

[1] We refer the reader to Chapter 2, Section 2.9 for a discussion of the distinction between prediction in the lay sense and in the sense of the Predictive Approach and prediction in the sense of statistical inference.

released from prisons. The data available to the study are limited in what they can reveal about the decision about whether or not to grant parole in the first place.[Bradley and Engen, 2016]

Chapter 3 describes the law and process of parole in California, with special attention to changes that have occurred during the period of our study. Existing studies of the parole decision process over those serving life sentences in California have been limited by the data available to each study. For example, a study that is comprehensive in the number of parole hearings included, are limited by the number of features available for analysis, is generally restricted to basic information such as the commitment offense and report basic demographics about the population of individuals serving life sentences.[Weisberg et al., 2011] Studies that require manual coding of additional features have been restricted to smaller subsets of hearings, such as a sample of 302 hearings from 2011 to understand the impact of Marsy's Law[Friedman and Robinson, 2014], a study of 754 hearings to understand the decision factors involved in the parole board granting parole[Young et al., 2015], or a study of criminalized masculinity among 109 hearings from 2017–2018. All studies rely on the Board of Parole Hearings to release a sample of hearings, which is released at the discretion of the Board and not necessarily random. A study of youth offender parole hearings included the complete set over an 18-month period and did not require sampling, but nonetheless is limited to 426 hearings.[Bell, 2019]

7.2 Data

Our dataset includes 34,993 hearing transcripts for which the decision outcome is known, covering 99.7% of all hearings. These transcripts feature 15,747 parole candidates. CDCR records four ethnic categories and two gender categories and in our dataset we observe 5,102 Black, 4,909 Latinx, 3,931 White, and 1,805 Other candidates. CDCR designates 15,014 of these individuals as male and 733 as female.

Factors for extraction and analysis were selected through deliberations with legal experts in parole, formerly incarcerated individuals, advocacy groups including appellate attorneys, representatives from the California Governor's office, as well as BPH[2]. The list of factors also built on previous work identifying relevant features including a feature set used to analyze a subset of 754 hearings from October 2007 – January 2010 [Young et al., 2015], also adapted for the analysis of a subset of 426 youth offender hearings from January 2014 – June 2015 [Bell, 2019].

A traditional study design requires a team of research assistants to read transcripts and manually record factors. To validate our methodology and findings, we replicated this setup and labeled 55 factors across a sample of 688 hearings (stratified by year).

Through the NLP efforts described in Chapter 6, we were able to reliably extract, for all 34,993 hearings, 18 of these factors at near human-level accuracy through an NLP pipeline. These factors

[2]Discussions with BPH included two conversations with Director of the Board of Parole Hearings Jennifer Shaffer in late 2018 and early 2019.

can only be obtained through computational methods, since no researcher could reasonably tabulate linguistic statistics for 150-page hearings. Following a court order [Superior Court of California in and for the County of San Francisco, 2020b], CDCR also provided 8 additional tabular factors.

Table 7.1: Legend for each feature contained in Table 7.2.

Feature	Description
Hearing Actors	
`retained attorney`	Whether the candidate privately engaged an attorney
`commissioner rate`*	Historical grant rate of the presiding commissioner at the time of the hearing
`victim oppose`	Does the victim make a statement opposing parole? (Used in the manual regression but not the NLP regression.)
`victim present`	Victim present at hearing? (Used in the NLP regression in place of `victim oppose`.)
`district attny oppose`	Does the DA make a statement opposing parole? (Used in the manual regression but not the NLP regression.)
`district attny present`	DA present at hearing? (Used in the NLP regression in place of `district attny oppose`.)
`attorney opinion`	In the closing statement did the candidate's attorney argue for release?
Time & Place	
`initial hearing`	Is this the candidate's first hearing?
`years since 2007`	Year of the hearing (since the first year of the dataset)
`years since eligible`	Number of years candidate has served over their lowest applicable parole eligibility date
Demographics	
`ethnicity black`	CDCR-recorded ethnicity = "Black"
`ethnicity latinx`	CDCR-recorded ethnicity = "Hispanic/Latino"
`ethnicity other`	CDCR-recorded ethnicity = "Other"
`gender female`	CDCR-recorded gender = female (indicated by CDCR number beginning with the letter `W`)
Pre-Commitment	
`justice involved`	Did candidate have prior interaction with the criminal justice system? Combines `prior convictions binary`, `prior supervision`, and `precommit prison`

`precommit sex abuse`	Victim of sexual abuse prior to commitment offense?
`precommit gang`	Whether person was involved in gang activity prior to commitment offense
Commitment Offense	
`offense murder second`	At least one count of Murder 2 (187)
`offense murder attempt`	At least one count of Attempted Murder (664-187 or 217)
`offense sex`	At least one count of rape/sexual assault (261-269)
`offense other`	Other (non Murder 1 baseline) offense
`crime gang`	Was crime rooted in gang activity?
`crime drugs alcohol`	Was the candidate heavily using alcohol/drugs around the time of the offense?
`claim innocence`	Does the candidate claim innocence in the commitment offense?
Rehabilitation	
`tabe edu score`	Most recent TABE score (or grade level equivalent if no TABE), histogram-bucketed into 3
`chronos bucket`	Number of laudatory chronos received, histogram-bucketed into 3
`programming all`	Total number of programs participated in ≥ 4 (Used in the manual regression but not the NLP regression)
`programming gang`	Participated in gang programming (Used in the NLP regression, since it was the only programming variable that we reliably extracted; in the manual regression, this is included as part of `programming all`.)
`12steps program failed`	Whether candidate was asked about the 12 steps and did not give an adequate response
`mental illness`	History of diagnosed mental illness?
`mental treatment`	Currently receiving mental health treatment (medication or counseling)?
Disciplinary	
`count 115s`	Total count of 115s (disciplinary writeup forms)
`clean time`	Years since last disciplinary infraction (Form 115 in the NLP regression; Form 115 or prison conviction in the manual regression)
`num pris convict buc`	Number of convictions while in prison, histogram-bucketed 3
`prison is level iv`	Whether hearing took place at a prison where more than half of the population is level IV
Parole Preparation	

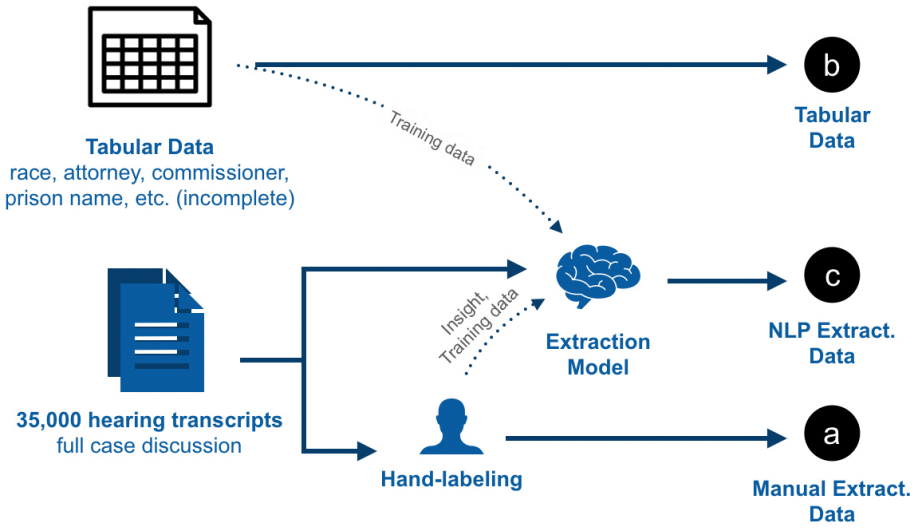

Figure 7.1: Data sources for primary regression analysis in Table 7.2. Factors for analysis were identified through discussions with legal experts and a broad set of stakeholders in parole. After hand-labeling a subset of factors and conducting an initial analysis, significant features were identified. This was followed by NLP extraction model development (using both human labels and weak supervision with labeling functions) and linguistic modeling.

`psych assess`	Psych Risk Score at most recent comprehensive assessment
`job offer`	Confirmed job offer?
Special Designation	
`youth offender`	Youth offender parole hearing - 3051/4801/260/261
`elderly parole`	Elderly Parole Designation

Table 7.1 describes each feature used in our analysis.

7.3 Methods

Our primary analysis model is a logistic regression onto the binary parole grant outcome using the NLP-extracted features. To validate the choice of this model and ensure confidence in its descriptive results and the insights they provide into the parole process, we construct two additional regression models:

(A) A regression that follows the traditional methodology of analyzing only the set of 688 manually coded transcripts. This regression aims to cover the broadest set of factors that are discussed in the course of a parole hearing. A set of 55 manually extracted factors are reduced and

bucketed into the 35 factors shown in Table 7.2, selected to remove strong collinearities and allow for easier interpretation while preserving all information.

(B) A regression that includes all 34,993 hearings but only analyzes the limited tabular labels provided directly by CDCR as well as fields parsed from the hearing title page.

We call the primary model with NLP-extracted labels regression C. Regression B provides a comparison baseline against which regressions A and C can be interpreted, where regression B is free from noise due to our human or automatic extraction. Under the non-differential error assumption, both models A and C underestimate significance slightly without explicit error-in-variable corrections [Gustafson, 2003]. For all three models, we validate our variable inclusion choices with robustness checks in Appendix 7.5. For each model separately, variables were chosen to avoid strong collinearities. However, ethnicity and gender were included as essential demographic factors for a descriptive analysis despite known collinearities. We analyze the adjusted odds ratio (AOR) $\exp(\beta)$ of each factor to provide a descriptive picture of how the factors jointly predict the parole decision. We assess statistical significance via Wald tests; AORs and p-values for these tests are reported in Table 7.2 across the three models.

7.4 Results

Can NLP help model parole decisions? We first assess the predictive fit of our models to validate the explanatory power that we can obtain through automatic extraction. We calculate the AUC statistic (Area Under the receiving operating Curve) under 10-fold cross validation [Stone, 1974, Geisser, 1975] for each model, shown in the second row of Table 7.2, which provides some model assessment guarantees [Debruyne et al., 2008, Liu et al., 2014, Beirami et al., 2017, Giordano et al., 2019, Wilson et al., 2020, Rad and Maleki, 2020]. The manual regression A attains an AUC of 0.804. The tabular regression B achieves an AUC of 0.730. Even though the tabular regression has access to a wealth of historical information, the small set of factors available in tabular form fails to explain parole hearing outcomes in the way that the dense manual regression with many fewer datapoints can. The NLP regression C achieves the highest AUC of the three models, 0.822. Including the automatically extracted variables provides a 9-point AUC boost over the tabular regression. While its feature set is less comprehensive than that of the manual regression, the massive sample size of the regression with NLP-extracted features enables it to provide the most accurate model of parole and more powerful estimates of the factor coefficients.

Are the NLP-based findings consistent with the traditional approach? Having established that the NLP-based model provides the most accurate descriptive picture of the parole process, we proceed to compare the consistency of the results across the three models on an item-by-item level. We observe that AORs overlap substantially for the factors identified as significantly predictive in the

Table 7.2: Regressions on the parole outcome based on (a) manually coded factors, (b) tabular data, (c) automatically extracted factors. Adjusted odds ratios (AORs) and Wald p-values are reported for all factors in parentheses. Significant values bold at $p < 0.05$.
*For historical commissioner grant rate, AOR is reported in units of a 10% increase in grant rate. For a description of each feature, see Table 7.1.

Data Source		(a) Manual	(b) Tabular	(c) NLP
n (Number of Hearings)		688	34,993	34,993
10-fold AUC		0.804	0.730	0.822
		Adjusted Odds Ratio e^{β} (p)		
Hearing Actors	`retained attorney`	1.71 (0.06)	**2.48 (0.00)**	**2.11 (0.00)**
	`commissioner rate`*	**1.74 (0.00)**	**1.34 (0.00)**	**1.41 (0.00)**
	`victim oppose`	**0.31 (0.00)**	-	-
	`victim present`	-	-	**0.42 (0.00)**
	`district attny oppose`	**0.26 (0.00)**	-	-
	`district attny present`	-	-	**0.69 (0.00)**
	`attorney opinion`	0.57 (0.10)	-	-
Time & Place	`initial hearing`	0.65 (0.24)	**0.40 (0.00)**	**0.45 (0.00)**
	`years since 2007`	1.11 (0.06)	**1.11 (0.00)**	**1.16 (0.00)**
	`years since eligible`	1.01 (0.75)	-	1.00 (0.05)
	`prison is level iv`	0.85 (0.73)	**0.32 (0.00)**	**0.57 (0.00)**
Demographics	`ethnicity black`	0.83 (0.55)	0.95 (0.17)	0.95 (0.23)
	`ethnicity latinx`	0.64 (0.18)	**1.14 (0.00)**	**0.91 (0.02)**
	`ethnicity other`	0.61 (0.22)	**1.24 (0.00)**	0.99 (0.78)
	`gender female`	0.92 (0.88)	**1.23 (0.00)**	**1.28 (0.00)**
Pre-Commitment	`justice involved`	**1.76 (0.02)**	-	-
	`precommit sex abuse`	0.80 (0.57)	-	-
	`precommit gang`	1.22 (0.58)	-	**1.25 (0.00)**
Commitment Offense	`offense murder second`	0.85 (0.53)	-	**1.13 (0.00)**
	`offense murder attempt`	1.06 (0.88)	-	**1.12 (0.03)**
	`offense sex`	0.34 (0.21)	-	**0.31 (0.02)**
	`offense other`	1.04 (0.92)	-	1.02 (0.60)
	`crime gang`	1.04 (0.92)	-	-
	`crime drugs alcohol`	1.04 (0.90)	-	-
	`claim innocence`	1.01 (0.98)	-	-
Programs & Rehabilitation	`tabe edu score`	1.15 (0.36)	-	**1.17 (0.00)**
	`chronos bucket`	**1.76 (0.01)**	-	-
	`programming gang`	-	-	**1.39 (0.00)**
	`programming all`	1.54 (0.18)	-	-
	`12steps program failed`	**0.31 (0.03)**	-	-
	`mental illness`	0.65 (0.06)	-	-
	`mental treatment`	1.22 (0.54)	-	-
Disciplinary	`count 115s`	1.00 (0.93)	-	-
	`clean time`	**1.06 (0.00)**	-	**1.02 (0.00)**
	`num pris convict buc`	0.77 (0.59)	-	-
Parole Preparation	`psych assess`	**0.49 (0.00)**	-	**0.48 (0.00)**
	`job offer`	**1.72 (0.04)**	-	**1.38 (0.00)**
Special Designation	`youth offender`	0.73 (0.39)	-	-
	`elderly parole`	0.79 (0.54)	-	-

NLP regression C and that all factors identified as significant in the NLP regression attain significance in the tabular regression B. Significant case factors that further raise chances of a favorable parole outcome in the manual regression A which are not tracked by the NLP regression include receiving laudatory chronos from prison programs and having been involved with the criminal justice system prior to the commitment offense, which may be seen as mitigating circumstances for the crime particularly for youth offenders. A significant additional factor that is not captured by the NLP model and decreases the chances of a grant under the manual model is failing to correctly answer a question about the "12 Steps" program of Alcoholics Anonymous/Narcotics Anonymous.

Many case factors in the manual regression do not meet the threshold for statistical significance. This may of course be attributed to its smaller sample size. Factors that attain significance in the larger model but not the manual model include whether the hearing is an initial parole hearing, whether the prison is considered max-security, the original offense, educational score, the total amount of programming, attorney representation status (marginally significant in regression A), ethnicity, and hearing year. The amount of time a candidate has served in prison does not appear to influence the parole outcome across either regression that track the variable. The legislature introduced two special designations for parole candidates in recent years: "youth offenders" and "elderly parole." These are tracked in the manual regression and neither designation attains significance.[3]

The NLP-extracted regression uses victim and district attorney presence instead of victim and district attorney opposition, respectively, because the machine learning classifier for opposition was unable to outperform a baseline for presence due to the strong class imbalance—victim representatives and district attorneys almost always oppose parole if they attend a hearing. In the manual regression, we use the more granular variables indicating whether the district attorney or victim made a statement opposing parole. We see that the manual regression attains higher-magnitude coefficients for these variables, but the signs, order of magnitude, and significance level match.

Finally, candidate ethnicity is significantly predictive of the parole outcome as an isolated variable in regressions B and C with inconsistent coefficients. However, the impact of race and ethnicity in parole cannot be studied merely through an independent regression variable, since we know that the independence assumption does not hold. For example, while ethnicity does not attain significance in regression A, many of the case factors considered have strong collinearities with ethnicity. Table 7.4 reveals that ethnicity significantly predicts the psychological risk assessment score candidates receive, the most predictive case factor in our models across all models. The regression analysis does not conclude that race or any other variable is (or is not) causing parole to be granted or denied. The analysis does, however, illuminate structural patterns in parole decision-making that undoubtedly call for further research into the causal effect of race in parole.

[3]This effect prevails in a robustness checks that restricts the hearing years of the regression to the years after the legislative change went into effect.

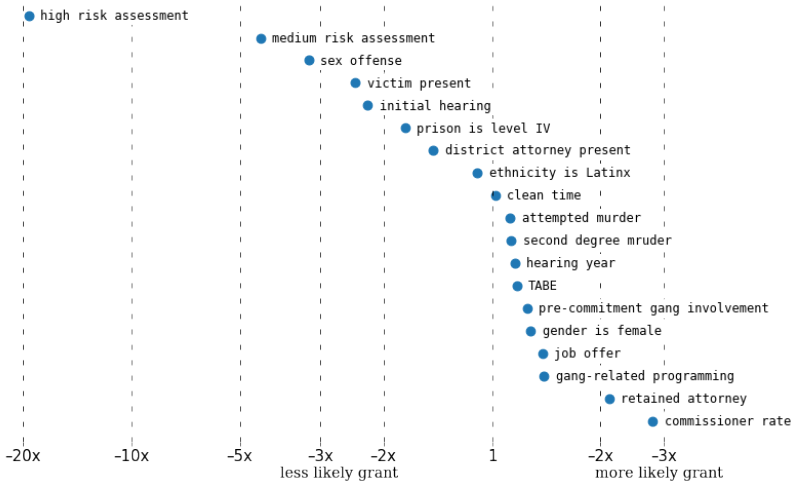

Figure 7.2: Adjusted odds ratios (AORs) for factors that achieve $p < 0.05$ significance in a Wald test in a regression modeling hearing outcome. For historical commissioner grant rate, AOR is reported in units of a 10% increase in grant rate. For a description of each feature, see Table 7.1. Table 7.2 shows complete regression coefficients.

Which factors most predict parole outcome? The factors in our analysis broadly fall into two categories: those that pertain to the parole case, such as a candidate's rehabilitational programming and disciplinary history, and factors which are generally outside of the candidate's control. Factors outside of candidates' control include demographics and factors that only become known at the time of the hearing, such as whether the district attorney is present, whether a victim representative appears, which year the hearing takes place, and which commissioner presides over the hearing.

In Figure 7.2, we plot the adjusted odds ratios for factors that attain significance in predicting parole outcomes. The past grant rate of the presiding commissioner administering a hearing significantly predicts the probability of a grant. A commissioner at the 10th percentile of grant rates has, at the time of the hearing, a historical grant rate of 4.76%. A commissioner at the 90th percentile of grant rates has a rate of 34.37%. Using only raw numbers, this is a ratio of 7.2x. Controlling for all other extracted factors, being assigned to a 90th percentile granting commissioner as opposed to a 10th percentile commissioner improves a candidate's chances of parole by 2.8x. The prison in which the parole candidate is held also significantly predicts the probability of a grant, motivating Study 2 below.

The dominating case factor in predicting the parole outcome in the model is the psychological risk assessment score. Under our model, candidates who score a "high" on the assessment are 19.4x less likely to receive a grant than candidates who score a "low," controlling for other factors. Similarly, candidates who score a "moderate" are 4.4x less likely to be released than candidates who score a "low." These differences can be explained in part by the fact that the psychological assessment

may account for some of the measured and unmeasured factors considered in a hearing, since it is conducted by a forensic psychologist with knowledge of various factors.[4] Nonetheless, the extent to which parole decisions align with psychological assessments is striking when compared to other case factors.

Examining other factors in the regression model, each additional year of `clean time` (the number of years since a candidate's last disciplinary infraction) improves the chances of a parole grant by only 2%. The amount of time a candidate has served in prison does not appear to predict the parole outcome. Candidates who retain an attorney are 2.2x as likely to receive a parole grant as candidates who are represented by a board-appointed attorney or elect to attend their hearings without an attorney. If a victim representative or the district attorney appears at the hearing (in almost all cases to make a statement opposing parole), each of them reduces the probability of a grant outcome by approximately half.

Significant case factors that further raise chances of a favorable parole outcome include a higher "TABE score," which measures a candidate's reading level grade equivalent, a pre-commitment gang affiliation, which may be seen as mitigating circumstances for the crime, and a parole plan: for example, having secured a job offer increases a candidate's chances of being granted parole by a factor of 1.4x. The nature of the original crime significantly impacts the chances of a favorable parole outcome, with second degree and attempted murder improving parole chances slightly compared to first degree murder[5], and sex offenders being 3.2x less likely to receive a parole grant.

Limitations

A descriptive regression analysis explains whether factors are predictive of the parole outcome, but not the mechanisms that underlie the relationship. One factor that illustrates this limitation is ethnicity. Controlling for the other case factors, Latinx candidates are half (0.55x) as likely to receive a parole grant as white candidates under the model. However, this adjusted odds ratio does not explain what other case factors mediate the effect of Latinx or other ethnicities. We hypothesize, but do not demonstrate, the existence of such mediators through variables that are significant in our regression. Research suggests that systemic racism plays a role in determining which candidates incur disciplinary infractions while in prison [Poole and Regoli, 1979, Heinz et al., 1976]. Racial bias has been documented in risk assessment [Angwin et al., 2016, Van Eijk, 2017, Mayson, 2018, Arnold et al., 2021]; Table 7.4 reveals that ethnicity significantly predicts the psychological risk assessment score candidates receive in our parole data as well. Another potential mediator that has

[4]The psychological risk assessment may be postulated to mediate the effects of other case factors. To validate our model, we separately regress the assessment score onto the variables reasonably known to the psychologist to evaluate the residual impact of the risk assessment. See Appendix 7.5.

[5]First degree murder is the most common offense in our population and is therefore used as the baseline offense in the regression. The adjusted odds ratios are relative to first degree murder. Attempted and second degree murder are "lower" offenses here. About 13% of hearings fall into the "other" offense category, which include controlling offenses such as kidnapping, etc.

been less studied in existing literature is the role of participation in a hearing. Figure 8.5 shows that speaking times differ significantly by ethnicity. When we augment the regression with various linguistic measures of participation, we find many of them to contribute additional predictive power even after controlling for case factors. The specific coefficients are reported in Table 8.3. One final mediator is the legal representation afforded to parole candidates, which is the focus of Study 3 below. Our analysis illuminates structural patterns in parole decision-making that undoubtedly warrant further research, including careful consideration of the causal arc of the variables and how race influences each one of them.

7.5 Regression Validation

7.5.1 Model Setup

We use a logistic regression model, where the dependent variable is the binary hearing outcome – a grant of parole, or a denial. Logistic regression takes on the following form:

$$\Pr[Y] = \sigma(\beta_0 + \beta^T x),$$

where $Y = 0$ refers to a denial of parole of any length, and $Y = 1$ refers to a grant of parole, $\sigma(t) = \frac{1}{1+e^{-t}}$ is the standard logistic function, $x \in \mathbb{R}^n$ is a representation of n features, and $\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^n$ are the parameters estimated. We interpret each β_j in the model as the additive effect on the log-odds of Y for each unit increase in x_j. We assess the significance of the coefficients identified by the model using Wald tests.

Since a few missing entries remain (even after the feature transformations detailed in Chapter 4, in the case of Table 7.2) and because our set of variables is large, we impute missing values with the column means. While this choice slightly underestimates the variance in the variables, it ensures that we do not overestimate the significance or effect sizes indicated by the coefficients for individual variables.

7.5.2 Robustness Checks on Table 7.2

We perform several tests to ascertain our choice of variables and justify their inclusion in the primary model:

1. Because our analysis covers a large time period that has seen much change across the legislative standard and case law that is relevant to parole in California as well as two changes in administration, we re-run our analyses using only data extracted for hearings from 2014 onwards. The results are given in Table 7.3.

2. The psychological assessment may account for many of the factors considered in a hearing since it is conducted by a forensic psychologist with knowledge of the case factors. In Table 7.4, we run a separate regression onto the psychologist's "comprehensive risk assessment" score using only those variables reasonably known to the psychologist. We can then attempt to assess the residual impact of the psychologist's opinion by subtracting the regressed score from actual value of the `psych assess` feature. The results for these models are given in Table 7.5.

3. To validate our choice of summing and bucketing programming participation into the `prog bucket` variable, we re-run the manual regression specification with the `prog bucket` variable replaced by the individual programming variables `progang`, `progartfit`, `progedu`, `proggang`, `progoth`, `progparent`, `progphil`, `progrel`, `progsubst`, `progther`, `progvictim`, and `progvoc`. The results for this model is given in Table 7.6.

4. To validate our findings about commissioner variability, we replace the `commissioner rate` variable with 52 individual indicator variables corresponding to the 52 commissioners in our dataset who conducted at least 50 hearings. Each variable indicates that the specified commissioner has conducted the hearing. We find that several of the commissioner indicator variables are significantly predictive of the parole outcome, some with positive and some with negative coefficients, controlling for the same case factors. The results for this model are given in Table 7.8.

5. Similarly, to validate our choice of the `prison level iv` variable, we replace this variable with 40 indicator variables corresponding to the 40 different prisons covered by our dataset. We find that several of the prisons that we identified as housing more than 50% of its population at security level IV are significantly predictive of parole denials. The results for this model are given in Table 7.9.

6. In our manual labeling effort, we tracked whether the board considered confidential information in their decision through the `confidential information` field. As a check, we re-run the manual specification with all hearings excluded where the board stated that they considered confidential information without explicitly stating that they did not rely on the information to arrive at the decision. The results are given in Table 7.11.

7. To better understand the effects of distributional drift that may be caused by our sampling of manually annotated hearings, we run formulations (a) and (c) only on the 688 manually labeled documents. The results are given in Table 7.12.

8. Because a candidates education level is collinear with their TABE score, our primary regression does not include the `edu level` feature. We repeat our analysis using `edu level` in place of TABE. Table 7.13 contains the results.

Table 7.3: Robustness check of Table 7.2. The hearings included are only those held in 2014 or later, which is when the youth offender and elderly parole designations started being used.

Data Source	(a) Manual	(b) Tabular	(c) NLP
n (Number of Hearings)	451	17,287	17,287
	Adjusted Odds Ratio e^{β} (p)		
retained attorney	1.54 (0.20)	**2.25 (0.00)**	**1.95 (0.00)**
initial hearing	0.65 (0.26)	**0.51 (0.00)**	**0.59 (0.00)**
years since 2007	**0.78 (0.02)**	**1.06 (0.00)**	**1.09 (0.00)**
ethnicity black	0.63 (0.22)	0.98 (0.63)	0.92 (0.11)
ethnicity latinx	0.88 (0.74)	**1.21 (0.00)**	0.98 (0.72)
ethnicity other	0.91 (0.84)	**1.34 (0.00)**	1.08 (0.23)
gender female	0.91 (0.87)	**1.30 (0.00)**	**1.39 (0.00)**
commissioner rate	**1.63 (0.01)**	**1.27 (0.00)**	**1.34 (0.00)**
prison is level iv	0.68 (0.46)	**0.33 (0.00)**	**0.57 (0.00)**
offense murder second	0.84 (0.57)	-	0.93 (0.18)
offense murder attempt	0.97 (0.94)	-	1.10 (0.14)
offense sex	**0.18 (0.05)**	-	**0.29 (0.01)**
offense other	1.79 (0.14)	-	0.96 (0.45)
years since eligible	0.99 (0.53)	-	**1.01 (0.00)**
precommit gang	0.74 (0.54)	-	**1.29 (0.00)**
tabe edu score	1.27 (0.16)	-	**1.14 (0.00)**
psych assess	**0.41 (0.00)**	-	**0.48 (0.00)**
clean time	**1.08 (0.00)**	-	1.00 (0.13)
job offer	1.48 (0.24)	-	**1.26 (0.00)**
programming gang	-	-	**1.33 (0.00)**
programming all	1.36 (0.42)	-	-
12steps program failed	0.34 (0.11)	-	-
victim oppose	**0.23 (0.00)**	-	-
victim present	-	-	**0.41 (0.00)**
district attny oppose	**0.28 (0.00)**	-	-
district attny present	-	-	**0.75 (0.00)**
youth offender	1.67 (0.20)	-	-
elderly parole	1.39 (0.44)	-	-
crime gang	2.05 (0.20)	-	-
crime drugs alcohol	0.54 (0.11)	-	-
precommit sex abuse	0.64 (0.39)	-	-
justice involved	**2.70 (0.00)**	-	-
num pris convict buc	0.76 (0.61)	-	-
mental illness	0.78 (0.43)	-	-
mental treatment	1.40 (0.39)	-	-
count 115s	1.01 (0.64)	-	-
chronos bucket	**2.16 (0.01)**	-	-
attorney opinion	1.16 (0.69)	-	-
claim innocence	1.16 (0.81)	-	-

Table 7.4: Regressions onto psychological risk assessment outcome based on the subset of factors that are reasonably determined at the time of the assessment.

Data Source	Manual	Tabular	NLP
n (Number of Transcripts)	688	34,993	34,993
	Adjusted Odds Ratio e^{β} (p)		
`initial hearing`	1.26 (0.06)	**1.48 (0.00)**	**1.44 (0.00)**
`prison is level iv`	**1.72 (0.00)**	**2.68 (0.00)**	**2.46 (0.00)**
`years since 2007`	1.03 (0.07)	**1.04 (0.00)**	**1.03 (0.00)**
`ethnicity black`	0.89 (0.30)	**1.24 (0.00)**	**1.18 (0.00)**
`ethnicity latinx`	0.84 (0.16)	1.01 (0.68)	1.02 (0.20)
`ethnicity other`	**0.75 (0.05)**	**0.84 (0.00)**	**0.84 (0.00)**
`gender female`	0.92 (0.64)	**0.71 (0.00)**	**0.70 (0.00)**
`offense murder second`	1.02 (0.87)	-	**0.91 (0.00)**
`offense murder attempt`	1.30 (0.07)	-	**1.10 (0.00)**
`offense sex`	1.45 (0.09)	-	**1.52 (0.01)**
`offense other`	1.22 (0.14)	-	1.02 (0.39)
`years since eligible`	1.01 (0.10)	-	**1.00 (0.00)**
`precommit gang`	1.23 (0.12)	-	**1.09 (0.00)**
`tabe edu score`	0.95 (0.33)	-	**0.89 (0.00)**
`clean time`	**0.97 (0.00)**	-	**1.00 (0.00)**
`job offer`	0.96 (0.71)	-	**0.69 (0.00)**
`programming gang`	-	-	**0.88 (0.00)**
`programming all`	0.85 (0.11)	-	-
`youth offender`	0.78 (0.10)	-	-
`elderly parole`	0.86 (0.38)	-	-
`crime gang`	0.86 (0.36)	-	-
`crime drugs alcohol`	0.88 (0.23)	-	-
`claim innocence`	1.21 (0.17)	-	-
`justice involved`	**1.32 (0.00)**	-	-
`num pris convict buc`	1.09 (0.61)	-	-
`mental illness`	**1.32 (0.00)**	-	-
`mental treatment`	**1.69 (0.00)**	-	-
`count 115s`	**1.02 (0.00)**	-	-
`chronos`	**0.82 (0.02)**	-	-

Table 7.5: Robustness check of Table 7.2. Using the residual variable `psych resid` instead of `psych assess`. Note that regression (b) does not include this feature, so its results remain unaffected.

Data Source	(a) Manual	(c) NLP
n (Number of Hearings)	688	34,993
	Adjusted Odds Ratio e^{β} (p)	
`retained attorney`	**1.68 (0.04)**	**2.11 (0.00)**
`initial hearing`	**0.46 (0.02)**	**0.34 (0.00)**
`years since 2007`	1.10 (0.08)	**1.13 (0.00)**
`ethnicity black`	0.90 (0.72)	**0.84 (0.00)**
`ethnicity latinx`	0.88 (0.67)	**0.89 (0.01)**
`ethnicity other`	0.96 (0.92)	**1.12 (0.03)**
`gender female`	1.00 (1.00)	**1.67 (0.00)**
`commissioner rate`	**1.75 (0.00)**	**1.41 (0.00)**
`prison is level iv`	0.50 (0.10)	**0.29 (0.00)**
`offense murder second`	0.83 (0.45)	**1.21 (0.00)**
`offense murder attempt`	0.65 (0.26)	1.04 (0.43)
`offense sex`	**0.16 (0.03)**	**0.23 (0.00)**
`offense other`	1.17 (0.62)	1.01 (0.81)
`years since eligible`	0.99 (0.73)	1.00 (0.23)
`precommit gang`	0.91 (0.80)	**1.17 (0.00)**
`tabe edu score`	**1.33 (0.04)**	**1.28 (0.00)**
`psych resid`	**0.46 (0.00)**	**0.48 (0.00)**
`clean time`	**1.09 (0.00)**	**1.02 (0.00)**
`job offer`	**2.02 (0.01)**	**1.81 (0.00)**
`programming gang`	-	**1.53 (0.00)**
`programming all`	1.24 (0.46)	-
`12steps program failed`	**0.30 (0.02)**	-
`victim oppose`	**0.30 (0.00)**	-
`victim present`	-	**0.42 (0.00)**
`district attny oppose`	**0.26 (0.00)**	-
`district attny present`	-	**0.69 (0.00)**
`youth offender`	1.03 (0.93)	-
`elderly parole`	0.92 (0.84)	-
`crime gang`	1.48 (0.36)	-
`crime drugs alcohol`	0.63 (0.08)	-
`precommit sex abuse`	0.76 (0.48)	-
`justice involved`	1.17 (0.51)	-
`num pris convict buc`	0.64 (0.36)	-
`mental illness`	**0.58 (0.02)**	-
`mental treatment`	0.77 (0.41)	-
`count 115s`	0.98 (0.12)	-
`chronos bucket`	**1.82 (0.00)**	-
`attorney opinion`	0.87 (0.64)	-
`claim innocence`	0.78 (0.50)	-

Table 7.6: Regressions on the parole outcome using individual programming variables instead of the `prog bucket` variable.

Factor	Adjusted Odds Ratio e^{β} (p)
n (Number of Hearings)	688
`retained attorney`	**1.99 (0.02)**
`initial hearing`	0.69 (0.33)
`years since 2007`	1.04 (0.47)
`ethnicity black`	0.66 (0.21)
`ethnicity latinx`	0.50 (0.06)
`ethnicity other`	0.60 (0.23)
`gender female`	0.66 (0.53)
`commissioner rate`	**1.94 (0.00)**
`prison is level iv`	1.02 (0.97)
`offense murder second`	0.89 (0.67)
`offense murder attempt`	1.12 (0.79)
`offense sex`	0.42 (0.33)
`offense other`	1.03 (0.93)
`years since eligible`	1.02 (0.23)
`precommit gang`	1.03 (0.94)
`tabe edu score`	1.17 (0.36)
`psych assess`	**0.48 (0.00)**
`clean time`	**1.06 (0.01)**
`job offer`	**1.91 (0.02)**
`12steps program failed`	**0.28 (0.02)**
`victim oppose`	**0.28 (0.00)**
`district attny oppose`	**0.28 (0.00)**
`youth offender`	0.67 (0.32)
`elderly parole`	0.70 (0.38)
`crime gang`	0.95 (0.91)
`crime drugs alcohol`	1.04 (0.91)
`precommit sex abuse`	0.71 (0.41)
`justice involved`	**2.29 (0.00)**
`num pris convict buc`	0.70 (0.51)
`mental illness`	0.68 (0.13)
`mental treatment`	1.10 (0.79)
`count 115s`	1.00 (0.81)
`chronos bucket`	**1.74 (0.01)**
`attorney opinion`	0.56 (0.13)
`claim innocence`	1.02 (0.96)
`progang`	1.35 (0.30)
`progartfit`	**2.27 (0.02)**
`progedu`	0.95 (0.83)
`programming gang`	**2.41 (0.01)**
`progparent`	0.72 (0.29)
`progphil`	1.65 (0.12)
`progrel`	1.41 (0.18)
`progsubst`	0.83 (0.62)
`progther`	0.97 (0.92)
`progvictim`	**1.81 (0.03)**
`progvoc`	0.87 (0.69)
`progoth`	0.92 (0.78)

Table 7.7: Robustness check of Table 7.2. Rather than use `commissioner rate` to measure presiding commissioners, we use a fixed effect on the individual commissioners.

Data Source	(a) Manual	(b) Tabular	(c) NLP
n (Number of Hearings)	688	34,993	34,993
	Adjusted Odds Ratio e^{β} (p)		
`retained attorney`	**2.14 (0.02)**	**2.48 (0.00)**	**2.11 (0.00)**
`initial hearing`	0.52 (0.11)	**0.40 (0.00)**	**0.46 (0.00)**
`years since 2007`	**1.21 (0.00)**	**1.11 (0.00)**	**1.20 (0.00)**
`ethnicity black`	0.91 (0.78)	0.95 (0.17)	0.96 (0.27)
`ethnicity latinx`	0.73 (0.40)	**1.14 (0.00)**	**0.90 (0.01)**
`ethnicity other`	0.58 (0.22)	**1.24 (0.00)**	0.98 (0.73)
`gender female`	1.00 (1.00)	**1.23 (0.00)**	**1.29 (0.00)**
`prison is level iv`	1.24 (0.67)	**0.32 (0.00)**	**0.55 (0.00)**
`offense murder second`	0.92 (0.79)	-	**1.13 (0.00)**
`offense murder attempt`	1.04 (0.93)	-	**1.13 (0.02)**
`offense sex`	0.38 (0.29)	-	**0.31 (0.02)**
`offense other`	1.18 (0.67)	-	1.05 (0.26)
`years since eligible`	1.01 (0.76)	-	1.00 (0.05)
`precommit gang`	1.10 (0.80)	-	**1.25 (0.00)**
`tabe edu score`	1.11 (0.55)	-	**1.16 (0.00)**
`psych assess`	**0.42 (0.00)**	-	**0.47 (0.00)**
`clean time`	**1.07 (0.00)**	-	**1.02 (0.00)**
`job offer`	1.65 (0.08)	-	**1.37 (0.00)**
`programming gang`	**-**	-	**1.39 (0.00)**
`programming all`	1.53 (0.22)	-	-
`12steps program failed`	0.34 (0.06)	-	-
`victim oppose`	**0.24 (0.00)**	-	-
`victim present`	**-**	-	**0.42 (0.00)**
`district attny oppose`	**0.23 (0.00)**	-	-
`district attny present`	**-**	-	**0.68 (0.00)**
`youth offender`	0.79 (0.56)	-	-
`elderly parole`	0.95 (0.91)	-	-
`crime gang`	1.22 (0.67)	-	-
`crime drugs alcohol`	0.94 (0.86)	-	-
`precommit sex abuse`	0.85 (0.71)	-	-
`justice involved`	**1.80 (0.03)**	-	-
`num pris convict buc`	0.98 (0.97)	-	-
`mental illness`	0.83 (0.46)	-	-
`mental treatment`	1.42 (0.32)	-	-
`count 115s`	1.00 (0.86)	-	-
`chronos bucket`	**1.84 (0.01)**	-	-
`attorney opinion`	0.55 (0.11)	-	-
`claim innocence`	1.18 (0.70)	-	-

Table 7.8: Continuation of Table 7.7. Reports the value of the fixed effect for each commissioner, whose names have been hidden.

Data Source	(a) Manual	(b) Tabular	(c) NLP
Fixed Effect	Adjusted Odds Ratio e^{β} (p)		
anonymous commissioner	0.13 (0.06)	**0.71 (0.00)**	**0.63 (0.00)**
anonymous commissioner	0.44 (0.68)	1.36 (0.14)	1.40 (0.14)
anonymous commissioner	**0.06 (0.04)**	**0.24 (0.00)**	**0.19 (0.00)**
anonymous commissioner	**0.05 (0.03)**	**0.25 (0.00)**	**0.21 (0.00)**
anonymous commissioner	**0.19 (0.02)**	**0.68 (0.00)**	**0.62 (0.00)**
anonymous commissioner	**0.05 (0.00)**	**0.79 (0.01)**	**0.78 (0.02)**
anonymous commissioner	**0.18 (0.03)**	**0.70 (0.00)**	**0.68 (0.00)**
anonymous commissioner	0.18 (0.15)	1.03 (0.85)	1.08 (0.62)
anonymous commissioner	0.00 (1.00)	**0.15 (0.00)**	**0.13 (0.00)**
anonymous commissioner	0.39 (0.36)	**2.00 (0.00)**	**2.08 (0.00)**
anonymous commissioner	0.31 (0.31)	1.04 (0.74)	0.89 (0.41)
anonymous commissioner	0.32 (0.43)	**0.59 (0.01)**	**0.47 (0.00)**
anonymous commissioner	**0.05 (0.03)**	**0.36 (0.00)**	**0.31 (0.00)**
anonymous commissioner	**1.00 (0.00)**	0.62 (0.28)	0.89 (0.79)
anonymous commissioner	0.70 (0.78)	**0.76 (0.02)**	**0.75 (0.03)**
anonymous commissioner	0.00 (1.00)	**0.14 (0.00)**	**0.11 (0.00)**
anonymous commissioner	**1.00 (0.00)**	0.00 (0.94)	0.00 (0.94)
anonymous commissioner	0.00 (0.99)	**0.47 (0.00)**	**0.52 (0.01)**
anonymous commissioner	**0.02 (0.00)**	**0.62 (0.00)**	**0.49 (0.00)**
anonymous commissioner	0.00 (1.00)	1.14 (0.55)	1.26 (0.33)
anonymous commissioner	**0.12 (0.00)**	**0.73 (0.00)**	**0.62 (0.00)**
anonymous commissioner	**0.05 (0.02)**	**0.14 (0.00)**	**0.14 (0.00)**
anonymous commissioner	0.35 (0.32)	**0.60 (0.00)**	**0.51 (0.00)**
anonymous commissioner	0.00 (0.99)	**0.10 (0.00)**	**0.09 (0.00)**
anonymous commissioner	**0.10 (0.00)**	**0.65 (0.00)**	**0.56 (0.00)**
anonymous commissioner	**0.02 (0.00)**	**0.38 (0.00)**	**0.28 (0.00)**
anonymous commissioner	0.00 (0.99)	**0.51 (0.00)**	**0.41 (0.00)**
anonymous commissioner	0.14 (0.15)	**0.16 (0.00)**	**0.15 (0.00)**
anonymous commissioner	**0.17 (0.03)**	**0.59 (0.00)**	**0.48 (0.00)**
anonymous commissioner	0.00 (0.99)	**0.51 (0.00)**	**0.48 (0.00)**
anonymous commissioner	0.10 (0.06)	**0.67 (0.00)**	**0.59 (0.00)**
anonymous commissioner	**0.06 (0.00)**	**0.59 (0.00)**	**0.49 (0.00)**
anonymous commissioner	**0.04 (0.00)**	**0.49 (0.00)**	**0.42 (0.00)**
anonymous commissioner	0.09 (0.09)	0.85 (0.28)	0.82 (0.23)
anonymous commissioner	**0.10 (0.01)**	**0.71 (0.00)**	**0.67 (0.00)**
anonymous commissioner	**0.16 (0.02)**	0.88 (0.16)	**0.78 (0.01)**
anonymous commissioner	0.00 (1.00)	**0.44 (0.01)**	**0.46 (0.02)**
anonymous commissioner	**0.06 (0.01)**	**0.47 (0.00)**	**0.39 (0.00)**
anonymous commissioner	3.13 (0.52)	0.77 (0.16)	0.75 (0.16)
anonymous commissioner	0.12 (0.08)	0.94 (0.63)	0.79 (0.10)
anonymous commissioner	0.00 (0.99)	**0.55 (0.00)**	**0.46 (0.00)**
anonymous commissioner	0.13 (0.17)	**0.41 (0.00)**	**0.33 (0.00)**
anonymous commissioner	**0.04 (0.02)**	**0.65 (0.00)**	**0.52 (0.00)**
anonymous commissioner	0.00 (0.99)	**0.25 (0.00)**	**0.21 (0.00)**
anonymous commissioner	0.00 (0.99)	**0.48 (0.00)**	**0.33 (0.00)**
anonymous commissioner	0.00 (0.99)	**0.14 (0.00)**	**0.14 (0.00)**
anonymous commissioner	0.00 (1.00)	0.81 (0.30)	0.69 (0.09)
anonymous commissioner	0.26 (0.07)	1.04 (0.63)	1.02 (0.82)
anonymous commissioner	0.00 (1.00)	0.73 (0.18)	0.68 (0.13)
anonymous commissioner	0.21 (0.36)	**0.33 (0.00)**	**0.28 (0.00)**
anonymous commissioner	0.29 (0.29)	**0.57 (0.00)**	**0.50 (0.00)**
anonymous commissioner	**0.06 (0.03)**	**0.70 (0.00)**	**0.58 (0.00)**
anonymous commissioner	-	0.00 (0.93)	0.00 (0.93)
anonymous commissioner	-	0.00 (0.98)	0.00 (0.98)
anonymous commissioner	-	0.00 (0.98)	0.00 (0.98)
anonymous commissioner	-	0.00 (0.95)	0.00 (0.95)
anonymous commissioner	-	0.00 (0.98)	0.00 (0.98)
anonymous commissioner	-	0.00 (0.96)	0.00 (0.95)
anonymous commissioner	-	0.00 (0.98)	0.00 (0.98)
anonymous commissioner	-	0.00 (0.98)	0.00 (0.98)
anonymous commissioner	-	0.64 (0.34)	0.55 (0.25)
anonymous commissioner	-	**0.12 (0.04)**	**0.13 (0.05)**
anonymous commissioner	-	0.00 (0.97)	0.00 (0.96)

Table 7.9: Robustness check of Table 7.2. Rather than use `prison is level iiv` to measure prisons, we use a fixed effect on the individual prisons.

Data Source	(a) Manual	(b) Tabular	(c) NLP
n (Number of Hearings)	688	34,993	34,993
	Adjusted Odds Ratio e^{β} (p)		
`retained attorney`	1.66 (0.09)	**2.49 (0.00)**	**2.13 (0.00)**
`initial hearing`	0.68 (0.32)	**0.40 (0.00)**	**0.45 (0.00)**
`years since 2007`	**1.13 (0.04)**	**1.12 (0.00)**	**1.16 (0.00)**
`ethnicity black`	0.73 (0.33)	**0.93 (0.04)**	0.94 (0.10)
`ethnicity latinx`	0.60 (0.16)	**1.09 (0.02)**	**0.89 (0.01)**
`ethnicity other`	0.58 (0.20)	**1.18 (0.00)**	0.96 (0.42)
`gender female`	0.73 (0.80)	1.13 (0.34)	1.04 (0.78)
`commissioner rate`	**1.66 (0.00)**	**1.32 (0.00)**	**1.40 (0.00)**
`offense murder second`	0.83 (0.52)	-	**1.12 (0.00)**
`offense murder attempt`	1.25 (0.60)	-	**1.12 (0.05)**
`offense sex`	0.24 (0.13)	-	**0.32 (0.02)**
`offense other`	1.02 (0.95)	-	1.02 (0.67)
`years since eligible`	1.00 (0.97)	-	1.00 (0.05)
`precommit gang`	1.35 (0.42)	-	**1.25 (0.00)**
`tabe edu score`	1.17 (0.35)	-	**1.17 (0.00)**
`psych assess`	**0.48 (0.00)**	-	**0.48 (0.00)**
`clean time`	**1.07 (0.00)**	-	**1.02 (0.00)**
`job offer`	**1.83 (0.03)**	-	**1.35 (0.00)**
`programming gang`	-	-	**1.40 (0.00)**
`programming all`	1.41 (0.33)	-	-
`12steps program failed`	**0.29 (0.03)**	-	-
`victim oppose`	**0.32 (0.00)**	-	-
`victim present`	-	-	**0.42 (0.00)**
`district attny oppose`	**0.23 (0.00)**	-	-
`district attny present`	-	-	**0.66 (0.00)**
`youth offender`	0.70 (0.36)	-	-
`elderly parole`	0.76 (0.51)	-	-
`crime gang`	1.17 (0.72)	-	-
`crime drugs alcohol`	1.08 (0.81)	-	-
`precommit sex abuse`	0.87 (0.74)	-	-
`justice involved`	**2.04 (0.01)**	-	-
`num pris convict buc`	1.14 (0.80)	-	-
`mental illness`	0.74 (0.24)	-	-
`mental treatment`	1.43 (0.30)	-	-
`count 115s`	1.01 (0.62)	-	-
`chronos bucket`	**1.87 (0.00)**	-	-
`attorney opinion`	0.64 (0.22)	-	-
`claim innocence`	1.03 (0.94)	-	-

Table 7.10: Continuation of Table 7.9. Reports the value of the fixed effect for each prison. SATF stands for Substance Abuse Treatment Facility. Eagle Mountain is short for Eagle Mountain Community Correctional Facility. Golden State is short for Golden State Medium Community Correctional Facility.

Data Source	(a) Manual	(b) Tabular	(c) NLP
Fixed Effect	Adjusted Odds Ratio e^{β} (p)		
California City Correctional Center	-	0.47 (0.33)	0.48 (0.36)
California Correctional Center	-	0.23 (0.16)	0.37 (0.37)
California Correctional Institution	0.34 (0.48)	**0.76 (0.02)**	0.90 (0.45)
California Health Care Facility	1.82 (0.48)	**0.71 (0.00)**	0.94 (0.55)
California Institution for Men	0.55 (0.64)	**0.67 (0.00)**	**0.74 (0.02)**
California Institution for Women	1.57 (0.77)	0.90 (0.48)	1.17 (0.33)
California Medical Facility	1.28 (0.72)	**0.59 (0.00)**	**0.77 (0.00)**
California Men's Colony	1.27 (0.67)	0.95 (0.46)	1.08 (0.26)
California Rehabilitation Center	-	2.11 (0.20)	1.61 (0.44)
California State Prison, Centinela	0.00 (1.00)	0.63 (0.10)	0.80 (0.44)
California State Prison, Corcoran	2.43 (0.23)	**0.37 (0.00)**	**0.65 (0.00)**
California State Prison, Los Angeles County	2.03 (0.44)	**0.60 (0.00)**	1.09 (0.49)
California State Prison, Sacramento	0.72 (0.81)	**0.25 (0.00)**	**0.55 (0.00)**
California State Prison, Solano	0.85 (0.75)	**0.83 (0.00)**	0.93 (0.24)
California SATF	1.66 (0.48)	**0.73 (0.00)**	**0.84 (0.03)**
Calipatria State Prison	0.00 (1.00)	**0.27 (0.00)**	**0.53 (0.00)**
Central California Women's Facility	1.65 (0.70)	1.02 (0.87)	1.28 (0.09)
Chuckawalla Valley State Prison	1.89 (0.28)	**1.38 (0.00)**	**1.39 (0.00)**
Correctional Training Facility	1.49 (0.43)	1.09 (0.17)	1.13 (0.06)
Deuel Vocational Institution	**6.83 (0.02)**	0.95 (0.64)	1.02 (0.85)
Eagle Mountain	-	4.52 (0.24)	**36.25 (0.01)**
Folsom State Prison	1.19 (0.79)	**0.77 (0.00)**	**0.79 (0.01)**
Golden State	-	0.00 (0.94)	0.00 (0.94)
High Desert State Prison	0.00 (0.99)	**0.18 (0.00)**	**0.39 (0.00)**
Ironwood State Prison	0.16 (0.22)	**0.80 (0.05)**	0.96 (0.73)
Kern Valley State Prison	0.00 (0.99)	**0.09 (0.00)**	**0.18 (0.00)**
Mule Creek State Prison	0.45 (0.35)	**0.64 (0.00)**	**0.69 (0.00)**
North Kern State Prison	0.00 (1.00)	**0.37 (0.02)**	0.56 (0.19)
Pelican Bay State Prison	0.00 (0.99)	**0.09 (0.00)**	**0.22 (0.00)**
Pleasant Valley State Prison	0.00 (0.99)	**0.51 (0.00)**	**0.67 (0.01)**
Richard J. Donovan Correctional Facility	0.64 (0.65)	**0.62 (0.00)**	0.93 (0.51)
Salinas Valley State Prison	0.00 (0.99)	**0.16 (0.00)**	**0.34 (0.00)**
San Quentin State Prison	3.10 (0.07)	0.97 (0.71)	0.99 (0.85)
Sierra Conservation Center	0.00 (1.00)	0.76 (0.14)	0.83 (0.36)
Valley State Prison	1.58 (0.60)	0.92 (0.35)	0.93 (0.46)
Ventura Youth Correctional Facility	-	0.00 (0.96)	0.00 (0.96)
Wasco State Prison - Reception Center	-	0.00 (0.95)	0.00 (0.96)

Table 7.11: Robustness check of Table 7.2. Regressions on the parole outcome for the manual specification, excluding cases where the board indicated that they considered confidential information without explicitly stating that they did not rely on the information to arrive at the decision.

Factor	Adjusted Odds Ratio e^{β} (p)
n (Number of Hearings)	569
`(Intercept)`	**0.11 (0.01)**
`retained attorney`	**1.88 (0.05)**
`initial hearing`	0.74 (0.46)
`years since 2007`	1.11 (0.09)
`ethnicity black`	0.90 (0.76)
`ethnicity latinx`	0.61 (0.19)
`ethnicity other`	0.79 (0.59)
`gender female`	0.80 (0.72)
`commissioner rate`	**1.62 (0.00)**
`prison is level iv`	0.67 (0.46)
`offense murder second`	0.81 (0.48)
`offense murder attempt`	1.17 (0.73)
`offense sex`	0.56 (0.50)
`offense other`	1.07 (0.87)
`years since eligible`	1.01 (0.57)
`precommit gang`	1.04 (0.92)
`tabe edu score`	1.10 (0.58)
`psych assess`	**0.52 (0.00)**
`clean time`	**1.07 (0.00)**
`job offer`	1.48 (0.17)
`programming all`	1.45 (0.30)
`12steps program failed`	0.37 (0.07)
`victim oppose`	**0.35 (0.00)**
`district attny oppose`	**0.23 (0.00)**
`youth offender`	0.64 (0.27)
`elderly parole`	0.61 (0.24)
`crime gang`	1.02 (0.97)
`crime drugs alcohol`	1.16 (0.68)
`precommit sex abuse`	0.64 (0.33)
`justice involved`	**1.95 (0.02)**
`num pris convict buc`	1.13 (0.81)
`mental illness`	0.68 (0.13)
`mental treatment`	1.09 (0.82)
`count 115s`	1.01 (0.43)
`chronos bucket`	**1.73 (0.02)**
`attorney opinion`	**0.45 (0.05)**
`claim innocence`	0.86 (0.71)

Table 7.12: Robustness check of Table 7.2. We run all three regression specifications over the 688 manually labeled documents, rather than all 34,993.

Data Source	Manual	Tabular	NLP
n (Number of Transcripts)	688	688	688
	Adjusted Odds Ratio e^{β} (p)		
`retained attorney`	1.71 (0.06)	**1.85 (0.01)**	1.63 (0.07)
`initial hearing`	0.65 (0.24)	**0.58 (0.03)**	0.64 (0.18)
`years since 2007`	1.11 (0.06)	1.04 (0.28)	1.04 (0.38)
`ethnicity black`	0.83 (0.55)	1.00 (0.99)	0.86 (0.60)
`ethnicity latinx`	0.64 (0.18)	0.89 (0.66)	0.62 (0.13)
`ethnicity other`	0.61 (0.22)	0.98 (0.95)	0.58 (0.15)
`gender female`	0.92 (0.88)	0.46 (0.11)	0.71 (0.54)
`commissioner rate`	**1.74 (0.00)**	**1.42 (0.00)**	**1.76 (0.00)**
`prison is level iv`	0.85 (0.73)	**0.42 (0.02)**	0.86 (0.72)
`offense murder second`	0.85 (0.53)	-	0.85 (0.51)
`offense murder attempt`	1.06 (0.88)	-	0.94 (0.87)
`offense sex`	0.34 (0.21)	-	0.24 (0.08)
`offense other`	1.04 (0.92)	-	1.06 (0.87)
`years since eligible`	1.01 (0.75)	-	1.00 (0.79)
`precommit gang`	1.22 (0.58)	-	1.00 (0.99)
`tabe edu score`	1.15 (0.36)	-	1.19 (0.25)
`psych assess`	**0.49 (0.00)**	-	**0.45 (0.00)**
`clean time`	**1.06 (0.00)**	-	**1.06 (0.00)**
`job offer`	**1.72 (0.04)**	-	**2.15 (0.00)**
`programming gang`	-	-	**2.41 (0.00)**
`programming all`	1.54 (0.18)	-	-
`12steps program failed`	**0.31 (0.03)**	-	-
`victim oppose`	**0.31 (0.00)**	-	-
`victim present`	-	-	**0.33 (0.00)**
`district attny oppose`	**0.26 (0.00)**	-	-
`district attny present`	-	-	0.81 (0.59)
`youth offender`	0.73 (0.39)	-	-
`elderly parole`	0.79 (0.54)	-	-
`crime gang`	1.04 (0.92)	-	-
`crime drugs alcohol`	1.04 (0.90)	-	-
`precommit sex abuse`	0.80 (0.57)	-	-
`justice involved`	**1.76 (0.02)**	-	-
`num pris convict buc`	0.77 (0.59)	-	-
`mental illness`	0.65 (0.06)	-	-
`mental treatment`	1.22 (0.54)	-	-
`count 115s`	1.00 (0.93)	-	-
`chronos bucket`	**1.76 (0.01)**	-	-
`attorney opinion`	0.57 (0.10)	-	-
`claim innocence`	1.01 (0.98)	-	-

Table 7.13: Robustness check of Table 7.2 using education level instead of the TABE score.

Data Source	Manual	Tabular	NLP
n (Number of Transcripts)	688	34,993	34,993
	Adjusted Odds Ratio e^{β} (p)		
retained attorney	1.70 (0.06)	**2.48 (0.00)**	**2.05 (0.00)**
initial hearing	0.69 (0.31)	**0.40 (0.00)**	**0.45 (0.00)**
years since 2007	1.11 (0.06)	**1.11 (0.00)**	**1.15 (0.00)**
ethnicity black	0.83 (0.54)	0.95 (0.17)	0.95 (0.16)
ethnicity latinx	0.64 (0.19)	**1.14 (0.00)**	0.95 (0.20)
ethnicity other	0.62 (0.22)	**1.24 (0.00)**	0.99 (0.89)
gender female	0.88 (0.82)	**1.23 (0.00)**	**1.25 (0.00)**
commissioner rate	**1.70 (0.00)**	**1.34 (0.00)**	**1.42 (0.00)**
prison is level iv	0.85 (0.73)	**0.32 (0.00)**	**0.58 (0.00)**
offense murder second	0.86 (0.58)	-	**1.14 (0.00)**
offense murder attempt	1.04 (0.92)	-	**1.12 (0.03)**
offense sex	0.29 (0.16)	-	**0.30 (0.01)**
offense other	0.99 (0.97)	-	1.03 (0.45)
years since eligible	1.00 (0.77)	-	**1.00 (0.05)**
precommit gang	1.29 (0.49)	-	**1.24 (0.00)**
education level	1.30 (0.06)	-	**1.21 (0.00)**
psych assess	**0.50 (0.00)**	-	**0.48 (0.00)**
clean time	**1.06 (0.00)**	-	**1.02 (0.00)**
job offer	**1.71 (0.04)**	-	**1.36 (0.00)**
programming gang	-	-	**1.38 (0.00)**
programming all	1.48 (0.23)	-	-
12steps program failed	**0.33 (0.03)**	-	-
victim oppose	**0.30 (0.00)**	-	-
victim present	-	-	**0.41 (0.00)**
district attny oppose	**0.26 (0.00)**	-	-
district attny present	-	-	**0.68 (0.00)**
youth offender	0.75 (0.41)	-	-
elderly parole	0.78 (0.52)	-	-
crime gang	1.06 (0.90)	-	-
crime drugs alcohol	1.09 (0.78)	-	-
precommit sex abuse	0.80 (0.57)	-	-
justice involved	**1.82 (0.02)**	-	-
num pris convict buc	0.81 (0.66)	-	-
mental illness	0.64 (0.05)	-	-
mental treatment	1.23 (0.52)	-	-
count 115s	1.00 (0.80)	-	-
chronos bucket	**1.73 (0.01)**	-	-
attorney opinion	0.59 (0.13)	-	-
claim innocence	1.00 (0.99)	-	-

7.6 Commissioner Variability in Granting Parole

Commissioner grant rates naturally vary depending on the prisons where they preside (different prisons house different populations) and the period they have served (due to various reforms, such as those described in Chapter 3). Consider two hypothetical commissioners with the same decision making process. If one commissioner primarily presides over hearings at California Men's Colony, which has minimum and medium security classes, and the other primarily presides over hearings at Pelican Bay State Prison, a "supermax" prison, we expect the two commissioners to have different empirical grant rates, even if there is no underlying difference in their decision-making. Similarly, a commissioner who served only in 2007–2008 would likely have a lower empirical grant rate than an identical commissioner who served in 2015–2019.

Commissioners are not assigned to hearings uniformly at random over all historical hearings. But, for a given prison in a given year, hearings can plausibly be considered to be assigned as-if random within that subset of commissioners. Rather than test the null hypothesis that all commissioner grant rates are equal across all prisons and years, we test the following null hypothesis: in a given prison and in a given year, grant rates are independent of the commissioner.

Using a Monte Carlo randomization inference [Abrams et al., 2012], we assess commissioner variability by testing the following null hypothesis: in a given year and in a given prison, grant rates are independent of the presiding commissioner.

That is, a commissioner's assignment to a hearing, with a fixed prison and year, does not depend on other case factors. For example, suppose that a parole candidate is scheduled for a hearing in June 2015, at some prison. We assume that this candidate may just as likely have been scheduled for May 2015; the fact that they are scheduled for June 2015 is not dependent on characteristics of the case, and therefore, that candidate's hearing is just as likely to have gone before a commissioner presiding over hearings at that prison in May, as before a commissioner presiding over hearings at that prison in June.

Therefore, any variation in grant rates conditioned on prison and year can be attributed to underlying commissioner variability. To test commissioner variability, we sample from this null distribution using Monte Carlo randomization inference.

In each sample, commissioner i does the same number of hearings at prison j in year k as in the original, historical data. It is only *which* hearings at prison j in year k that may differ. Under the null hypothesis, these are hearings that commissioner i may plausibly have presided over instead; they were still doing hearings at the "same place" (prison j) at the "same time" (year k). The only reason that commissioner i wasn't assigned them is due to random assignment. This null hypothesis considerably constricts how far any sample differs from the empirical data. Consider a prison j and year k in which only one hearing occurred. The same commissioner will be assigned to this hearing in every sample, which is also the same assignment as in the empirical data.

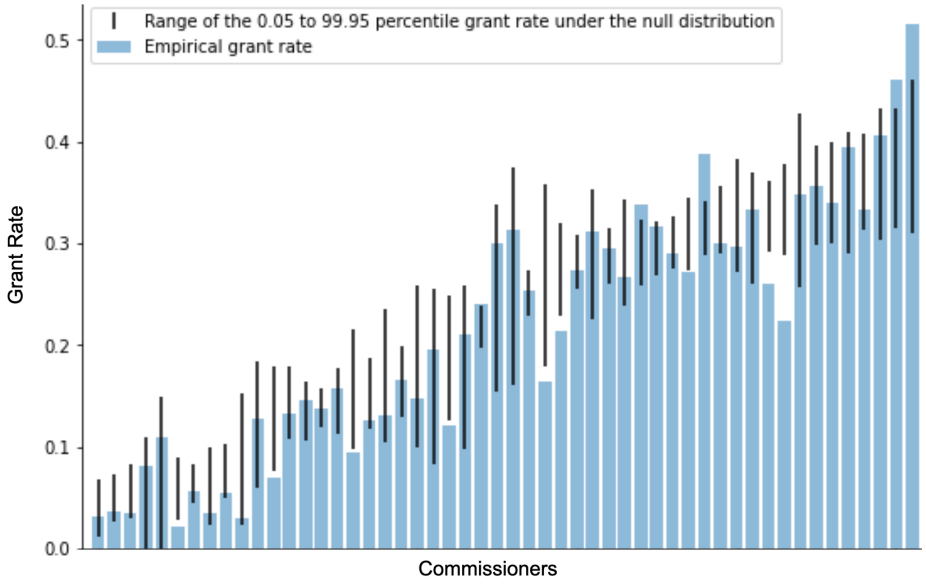

Figure 7.3: Grant rates of 52 individual commissioners. The blue bar shows the empirical grant rate of a commissioner. The black vertical bar shows the 0.05–99.95 percentile range of the grant rate for that commissioner under a null hypothesis where the grant rates are equal within a given prison in a given year, independent of presiding commissioner. The commissioners are sorted left to right in increasing order of the median null grant rate.

For each assignment sample, we calculate each commissioner's grant rate (excluding commissioners who have presided over fewer than 50 hearings during their tenure). We consider 10,000 assignments sampled under the null distribution for each of 52 commissioners and the resulting grant rates under this null distribution.

We reject the null hypothesis across commissioners at a family-wise error rate of $\alpha = 0.05$. We find that 14 of the 52 commissioners fall outside their Bonferoni-corrected intervals at this $\alpha = 0.05$ level, shown in Figure 7.3 as the 0.05–99.95 percentile range. That is, the 14 commissioners have higher or lower grant rates than expected assuming that prison and year are the sole sources of variation in a commissioner's grant rate.

Chapter 8

Linguistic Discrepancies Among Parole Attorneys

The right to legal counsel is guaranteed in a number of legal jurisdictions and contexts, such for criminal defendants in U.S. federal court through the Sixth Amendment of the U.S. Constitution.[1] In many such jurisdictions, the right is guaranteed through the availability of free civil counsel. However, the quality and legitimacy of the representation offered by such counsel has been called into question for reasons such as lack of funding, case overload, and conflict of interest [Ovalle, 2021].

Studying the quality of representation from a lexical perspective is crucial to understanding representation as component of procedural justice, but existing analyses of free counsel rely on measuring outcomes, such as incarceration rate or sentence length for public defenders [Hartley et al., 2010, Williams, 2013, Cohen, 2014], or on perceptions of public defenders [Posner and Yoon, 2010]. Such studies provide valuable insights into the effectiveness of counsel, but they are nonetheless indirect measures of attorneys' behavior.

We provide an analysis that is unique in two ways: first, we show the effectiveness of supervised and unsupervised text-based methods in understanding attorney representation, and second, we study attorneys in California's parole hearing system for individuals who are already incarcerated, unlike the vast majority of existing studies on the U.S., which focus on public defenders in the criminal trial system.

We study differences between board-appointed and retained attorneys in a corpus of 35,105 parole hearings from the State of California for individuals serving life sentences between 2007 and 2019. At a high level, the two types of attorneys differ in their speaking time, lexical complexity, and syntactic complexity. We further study their lexical differences using two methods. The first is unsupervised and identifies distinguishing words used by either type of attorney. The second measures differences

[1] A broad basis for a civil right to counsel is given in Article 10 of the International Declaration of Human Rights.

in a lexicon designed by parole scholars.

Finally, for all identified differences between board-appointed and retained attorneys, we show that they predict the outcome of the parole hearing, even when controlling for case factors.

This analysis is particularly relevant today as recent debate in the California legislature has concerned the effectiveness, compensation, and training of board-appointed attorneys [Committee on Revision of the Penal Code, 2020].

8.1 Data

We study discrepancies in attorney speech across the 34,993 parole hearing transcripts described in Chapter 4. Of the hearings, the candidate is represented by a board-appointed attorney in 6,825 of them, and by a retained attorney in 25,542. The attorney status is unknown in 2,626 hearings. Unless otherwise stated, the analyses in this chapter study the subset of 32,367 hearings in which attorney status is known.

8.2 Who Gets Retained Counsel?

Our data show substantial demographic discrepancies in attorney representation. For example, only 17.7% and 18.3% of hearings with Black and Latinx candidates, respectively, have a retained attorney, compared to 27.2% of hearings with White candidates. In other words, Black and Latinx candidates are two-thirds as likely as White candidates to be represented by a retained attorney. Similarly, 43.6% of hearings with female parole candidates have a retained attorney, compared to 20.0% of hearings with male candidates, making female candidates just more than twice as likely as male candidates to have a retained attorney.

To better understand how demographic factors and other hearing factors correlate with attorney representation, we perform logistic regressions to predict whether or not a candidate will obtain private legal representation. Similar to the methodology described in Section 7.3, we perform three regressions, where the first includes only the manually coded transcripts, the second includes all transcripts (with known attorney status) but only with tabular features provided by CDCR, and the third includes all transcripts (with known attorney status) with NLP-extracted features.

For full definitions and descriptions of all features, see Table 7.1. Compared to the analyses in Chapter 7, we restrict the analysis in the present chapter to the subset of factors that are reasonably known to the candidate and their support network at the time at which they decide whether to retain an attorney.

Table 8.1 presents adjusted odds ratios (AORs) and Wald p-values for each feature for each of the three regressions. For readability, we have plotted the AORs of all features that achieve $p < 0.05$ significance in the regression with NLP-extracted features. Case factors that make it more likely for

Table 8.1: Regressions onto attorney representation based on the subset of factors that are reasonably known to the candidate at the time they decide whether to retain an attorney, over the set of hearings where attorney representation status is known.

Data Source	(a) Manual	(b) Tabular	(c) NLP
n (Number of Transcripts)	623	32,349	32,349
	Adjusted Odds Ratio e^{β} (p)		
`initial hearing`	**1.84 (0.03)**	**0.91 (0.01)**	**0.91 (0.01)**
`prison is level iv`	0.51 (0.10)	**0.89 (0.01)**	0.98 (0.60)
`years since 2007`	0.95 (0.17)	**0.97 (0.00)**	**0.97 (0.00)**
`ethnicity black`	**0.46 (0.01)**	**0.61 (0.00)**	**0.59 (0.00)**
`ethnicity latinx`	**0.53 (0.03)**	**0.65 (0.00)**	**0.55 (0.00)**
`ethnicity other`	0.58 (0.11)	**0.90 (0.02)**	**0.81 (0.00)**
`gender female`	**2.77 (0.00)**	**2.86 (0.00)**	**3.40 (0.00)**
`offense murder second`	0.93 (0.75)	-	**0.92 (0.03)**
`offense murder attempt`	0.85 (0.65)	-	**0.90 (0.04)**
`offense sex`	0.41 (0.18)	-	**0.39 (0.04)**
`offense other`	0.78 (0.44)	-	**0.84 (0.00)**
`years since eligible`	**1.04 (0.01)**	-	1.00 (0.78)
`precommit gang`	0.81 (0.58)	-	**1.15 (0.00)**
`tabe edu score`	1.05 (0.74)	-	**1.22 (0.00)**
`clean time`	1.02 (0.31)	-	1.00 (0.97)
`job offer`	1.62 (0.06)	-	**2.07 (0.00)**
`programming gang`	-	-	**1.34 (0.00)**
`programming all`	1.07 (0.79)	-	-
`youth offender`	1.65 (0.15)	-	-
`elderly parole`	0.73 (0.46)	-	-
`crime gang`	1.61 (0.27)	-	-
`crime drugs alcohol`	**0.59 (0.04)**	-	-
`claim innocence`	**0.29 (0.01)**	-	-
`justice involved`	0.71 (0.13)	-	-
`num pris convict buc`	0.88 (0.78)	-	-
`mental illness`	1.21 (0.42)	-	-
`mental treatment`	0.88 (0.69)	-	-
`count 115s`	1.00 (0.75)	-	-
`chronos`	**1.75 (0.01)**	-	-

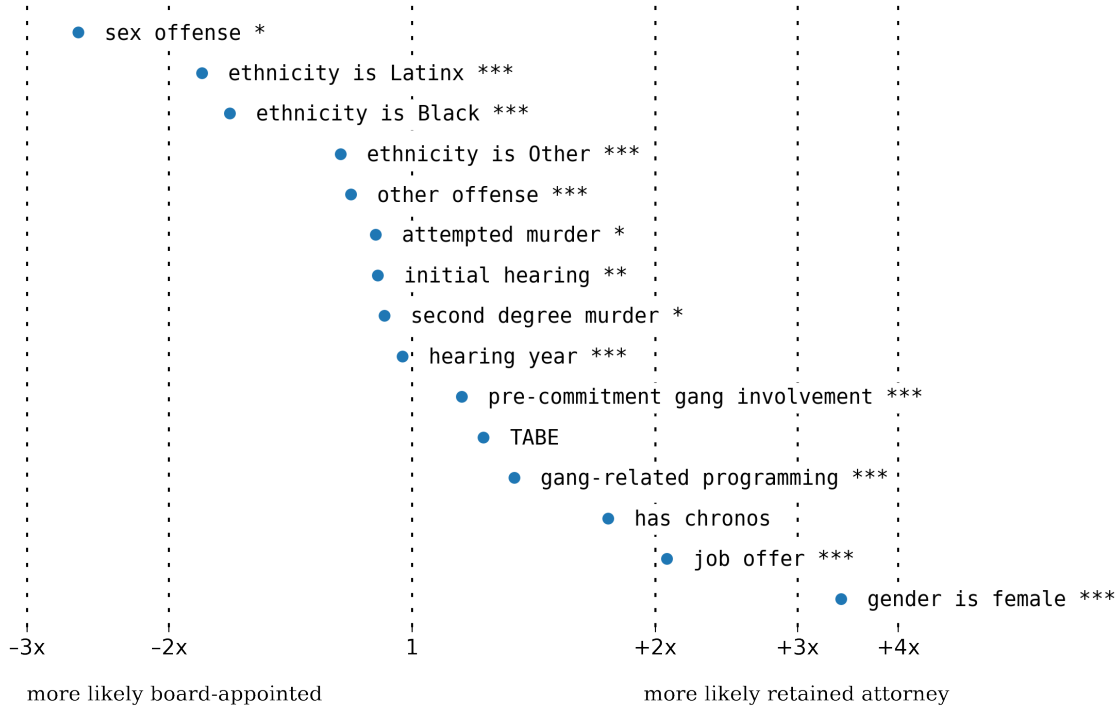

Figure 8.1: Adjusted odds ratios (AORs) for factors that achieve $p < 0.05$ significance in a Wald test in a regression modeling attorney status. For a description of each feature, see Table 7.1. Table 8.1 shows complete regression coefficients.

a candidate to retain counsel include having a confirmed job offer and laudatory chronos from prison staff. Candidates who claim to be innocent of their commitment offense are 3.4x less likely to retain an attorney. Female candidates are 3.4x more likely to hire a retained attorney than male candidates. As mentioned above, without controlling for other features, female candidates were instead found to be 2.1x more likely to be represented by a retained attorney. Black and Latinx candidates are approximately 0.5x as likely as white candidates to attend their hearing with a privately retained lawyer, controlling for the other factors.

8.3 Discrepancies in Attorney Language

In the following sections, we measure differences between board-appointed and retained attorneys across three sets of features. Each section defines the features used and the methods used to determine those features. In addition to describing the discrepancies, we also measure whether the discrepancies are predictive of the eventual parole outcome. The method for doing so is the same for each section. We use logistic regression to predict the binary outcome variable of the parole hearing (a grant or a denial of parole), based on a set of existing case factors associated with the dataset. In other words, we use the set of tabular and NLP-extracted features defined in Table 7.1, which are available for all hearings. In addition to the twenty-two existing tabular and NLP-extracted features, for each discrepancy we identify in the sections below (e.g. speaking time), we include the feature as a twenty-third feature in the logistic regression.

8.3.1 Speaking Time

Feature definitions

Procedural justice holds that the recipients of justice ought to have the opportunity to make their voices and perspectives heard [Tyler, 2003, Solum, 2004]. We now explore how the voices of the actors involved in a parole hearing influence the decision outcome.[2]

Three participants typically dominate the dialogue of a parole hearing: the candidate, their attorney, and the commissioner. We calculate three psycholinguistic measures of voice for each speaker. The first measure, speaking time (calculated as the absolute and relative number of words spoken), is a measure of quantity: "how much is said" by each participant. The other two measures are of quality and encode "how it is said." Lexical complexity is computed via the mean Age of Acquisition (AOA) statistic [Kuperman et al., 2012], referring to the age at which a word is typically learned. Syntactic complexity is computed through the mean sentence length [Szmrecsanyi, 2004]

[2]Parole hearings are procedural interrogations of the parole candidate through the commissioner. While victim representatives, district attorneys, and lawyers for the candidate are permitted to make closing statements, they may not converse directly with the candidate. Only the commissioners may direct questions to the candidate. Should the candidate's attorney wish to ask a question, the attorney must direct it to the commissioners who may then ask the candidate.

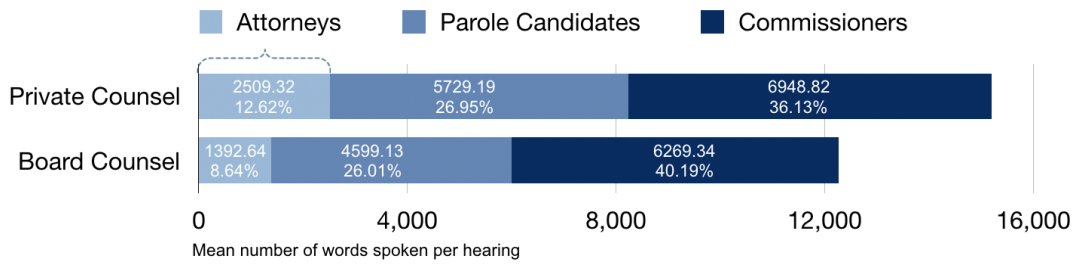

Figure 8.2: Speaking time of various individuals in a hearing, broken down by attorney status. All mean differences are statistically significant ($p < 0.05$) in a Chi-squared test.

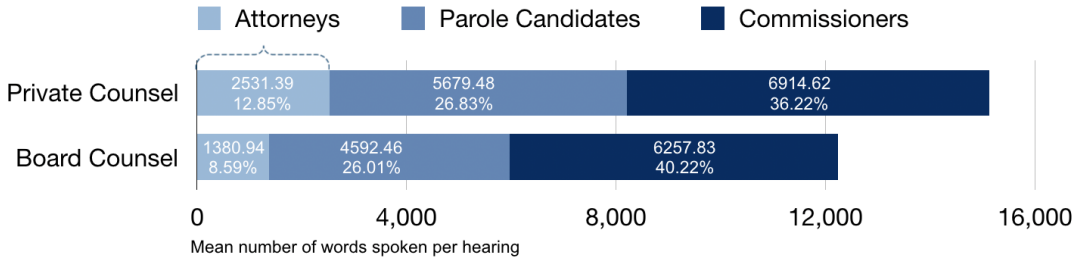

Figure 8.3: Speaking times for candidates, counsel, and commissioners conditioned on counsel status and restricted to hearings in which the candidate is identified as male.

of the speaker. Together, lexical complexity and syntactic complexity offer a view into the kinds of language used by a given speaker.

Fig. 8.2 shows that speaking times for the parole candidate, attorney, and commissioner are greater for hearings conducted with a private attorney vs. a board-appointed attorney. Private attorneys speak almost twice as many words as board-appointed attorneys and candidates speak 25% more in the presence of a privately retained attorney, with 18% longer hearings overall. These differences persist when accounting for candidate ethnicity and gender as shown in Figures 8.5, 8.3, and 8.4. We find that private attorneys speak with significantly higher lexical and syntactic complexity than board-appointed attorneys (Table 8.2). Table 8.5 shows that both speaking time and lexical and syntactic predict hearing outcome when controlling for other case factors.

We now investigate the breakdown of measures of voice in the population of parole candidates. Fig. 8.5 shows speaking proportions by ethnicity. All non-white candidate groups have lower hearing participation than white candidates. Latinx parole candidates and their lawyers speak fewer than

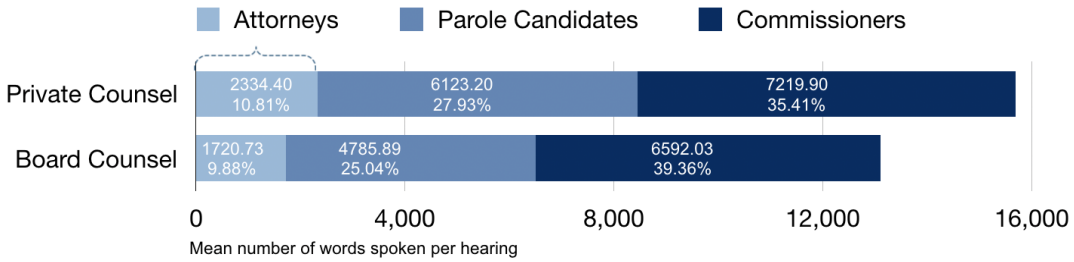

Figure 8.4: Speaking times for candidates, counsel, and commissioners conditioned on counsel status and restricted to hearings in which the candidate is identified as female.

Table 8.2: Average speaking time, lexical complexity, and syntactic complexity by attorney status.

Linguistic Factor	Privately Retained	Board-Appointed
`cand speaking raw`	5380.87	6606.69
`attn speaking raw`	1289.35	2285.47
`comm speaking raw`	4720.39	5218.42
`cand avg acq score`	4.82	4.86
`attn avg acq score`	5.41	5.47
`comm avg acq score`	5.19	5.19
`cand avg sent len`	9.23	9.95
`attn avg sent len`	14.90	15.20
`comm avg sent len`	9.68	9.74

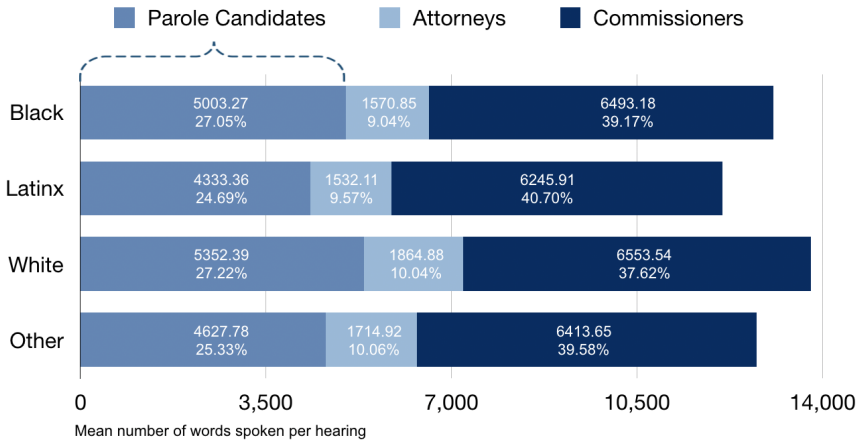

Figure 8.5: Speaking time by word count for parole hearings by parole candidate ethnicity (all differences statistically significant ($p < 0.05$) in t-tests). Latinx parole candidates speak on average 20% or 1,058 fewer words than white ones, which may predict their parole outcome as Table 8.3 shows.

Table 8.3: Regression on the parole outcome with linguistic markers of hearing participation. Each row is a rerun of the regression in Table 7.2 column (c) ($n = 34,993$) with the linguistic feature of inquiry added for the three speakers. (Therefore, the β coefficients are comparable across the columns but not the rows.) Each independent regression controls for all factors from Table 1(c). See Table 8.4 for all coefficients.

	Coefficient β (Wald test p)		
Factor Regression	Candidate	Attorney	Comm.
Speaking Time (10^5 words)	**2.35 (0.00)**	1.98 (0.18)	**-4.32 (0.00)**
Syntactic Complexity	**0.39 (0.00)**	**0.28 (0.00)**	**0.54 (0.00)**
Lexical Complexity	**0.06 (0.00)**	0.00 (0.16)	**-0.03 (0.00)**

white candidates and their lawyers, both proportionally, when compared to parole commissioners, and in absolute counts. White candidates speak an average of 24% or 1,058 more words than Latinx candidates in their hearings.

Regressions onto the parole hearing outcome

We analyze how hearing participation impacts the hearing outcome by regressing onto the parole outcome. We run three regressions, one for each measure. Each regression assesses the impact of one linguistic factor for the three speakers, controlled for all analysis factors in the NLP regression from Study 1 over 34,993 documents. The coefficients extracted from the three regressions are shown in the corresponding rows of Table 8.3. Controlling for case and hearing factors, all three measures significantly predict hearing outcomes for some participants. For both parole candidates and their attorneys, longer and more complex speech is more likely to lead to a grant in parole. The more the commissioner speaks, the more likely the hearing is to result in a denial. A candidate in the 10th percentile of participation speaks 1,592 words per hearing, while a candidate in the 90th percentile speaks 9,993 words. Other factors being equal, the candidate in the 90th percentile of speaking times is 22% more likely to be granted parole than the candidate in the 10th percentile. This effect suggests that one way in which race and ethnicity modulate parole outcomes is by determining whose voices get heard in the hearing procedure.

8.3.2 Word Polarity Analysis

We now investigate what private attorneys say that distinguishes them from board-appointed lawyers. We first conduct an exploratory word polarity analysis. Using the model-based word score method derived from normalized log odd ratios [**?**], we identify the most polar words that explain the difference between private and board attorney speech.

Table 8.4: Regressions on the parole outcome based on linguistic features of voice, controlling for case and hearing factors. Each column is a rerun of the regression in Table 7.2 column (c) with the linguistic feature of inquiry added for the three speakers. (Therefore, the β coefficients are comparable across the rows but not the columns.)

Linguistic Factor	Speak. Time	Lex. Complex.	Synt. Complex.
n (Number of Hearings)	34,993		
Speaking Time	Coefficient β (p)		
attn speaking raw	1.98 (0.18)	-	-
cand speaking raw	**2.35 (0.00)**	-	-
comm speaking raw	**-4.32 (0.00)**	-	-
Lexical Complexity			
attn lexical complexity	-	**0.28 (0.00)**	-
cand lexical complexity	-	**0.39 (0.00)**	-
comm lexical complexity	-	**0.54 (0.00)**	-
Syntactic Complexity			
attn syntactic complexity	-	-	0.01 (0.16)
cand syntactic complexity	-	-	**0.06 (0.00)**
comm syntactic complexity	-	-	**-0.03 (0.00)**
Control Variables			
retained attorney	**0.73 (0.00)**	**0.71 (0.00)**	**0.71 (0.00)**
initial hearing	**-0.79 (0.00)**	**-0.78 (0.00)**	**-0.80 (0.00)**
years since 2007	**0.15 (0.00)**	**0.15 (0.00)**	**0.14 (0.00)**
ethnicity black	-0.04 (0.33)	-0.03 (0.48)	-0.03 (0.41)
ethnicity latinx	-0.08 (0.07)	-0.04 (0.38)	-0.05 (0.22)
ethnicity other	0.00 (0.95)	0.03 (0.57)	0.03 (0.61)
gender female	**0.25 (0.00)**	**0.25 (0.00)**	**0.25 (0.00)**
commissioner rate	**0.33 (0.00)**	**0.34 (0.00)**	**0.33 (0.00)**
prison is level iv	**-0.55 (0.00)**	**-0.55 (0.00)**	**-0.56 (0.00)**
offense murder second	**0.12 (0.00)**	**0.13 (0.00)**	**0.13 (0.00)**
offense murder attempt	**0.11 (0.04)**	**0.12 (0.03)**	**0.12 (0.03)**
offense sex	**-1.15 (0.02)**	**-1.20 (0.01)**	**-1.14 (0.02)**
offense other	0.02 (0.68)	0.01 (0.74)	0.02 (0.59)
years since eligible	0.00 (0.05)	0.00 (0.05)	**0.00 (0.05)**
precommit gang	**0.22 (0.00)**	**0.21 (0.00)**	**0.20 (0.00)**
tabe edu score	**0.14 (0.00)**	**0.14 (0.00)**	**0.13 (0.00)**
psych assess	**-0.74 (0.00)**	**-0.73 (0.00)**	**-0.73 (0.00)**
clean time	**0.02 (0.00)**	**0.02 (0.00)**	**0.02 (0.00)**
job offer	**0.32 (0.00)**	**0.29 (0.00)**	**0.29 (0.00)**
programming gang	**0.32 (0.00)**	**0.32 (0.00)**	**0.31 (0.00)**
victim present	**-0.86 (0.00)**	**-0.88 (0.00)**	**-0.87 (0.00)**
district attny present	**-0.37 (0.00)**	**-0.39 (0.00)**	**-0.40 (0.00)**

Feature definitions

We model word usage as follows:

$$\mathbf{y} \sim \texttt{Multinomial}(n, \pi)$$

where $\mathbf{y}$ is the vector of term word counts for the entire corpus, n is the total number of words in the corpus, and π is the vector of probabilities for each word in the vocabulary. To account for inherent differences in word usage not based on the examined feature, the model is typically instantiated with a Dirichlet prior with parameter vector α, a vector of counts for each word in the corpus. Intuitively, α can be thought of as the number of times each word has been encountered *before* examining the corpus. For our experiments, we set α to be the vector of word counts across all attorney speech in all hearings, regardless of the attorney's status.

Given an observed vector of word counts from the corpus, y, the prior distribution, and the total number of words in the corpus n, the maximum likelihood estimate of the underlying probability distribution π is $\hat{\pi} = \frac{1}{n+\alpha_0} \cdot (y + \alpha)$, where α_0 is the sum of α_w for each word w in the corpus. We let a and b indicate the disjoint subsets of our corpus yielded by the feature under examination using superscripts, such that $y^{(a)}$ indicates the vector of word counts for that particular subset, with $\alpha^{(a)}$ and $n^{(a)}$ defined analogously. Under these specifications, we can estimate the odds of a specific word w compared to others for a subset a as $\hat{\Omega}_w^{(a)} = \frac{\hat{\pi}_w^{(a)}}{1-\hat{\pi}_w^{(a)}}$. From this, in turn, we can estimate the log-odds ratio for the word w between the two groups a and b (denoted $\hat{\delta}_w^{(a-b)}$) as follows:

$$\begin{aligned}\hat{\delta}_w^{(a-b)} = &\log \frac{(y_w^{(a)} + \alpha_w^{(a)})}{(n^{(a)} + \alpha_0^{(a)} - y_w^{(a)} - \alpha_w^{(a)})} \\ &- \log \frac{(y_w^{(b)} + \alpha_w^{(b)})}{(n^{(b)} + \alpha_0^{(b)} - y_w^{(b)} - \alpha_w^{(b)})}\end{aligned}$$

One of the important benefits of using a model-based approach (as opposed to just computing the log odds ratio directly from the vector of word counts), is that it offers not just a score for each word, but an estimate of the variance for that score. In particular, the variance is estimated as:

$$\hat{\sigma}^2(\hat{\delta}_w^{(a-b)}) = \frac{1}{(y_w^{(a)} + \alpha_w^{(a)})} + \frac{1}{(y_w^{(b)} + \alpha_w^{(b)})}.$$

So, instead of reporting the raw scores for any given word, we can instead report the normalized *z-score*, defined as:

$$z_w^{(a-b)} = \hat{\delta}_w^{(a-b)} / \sqrt{\hat{\sigma}^2(\hat{\delta}_w^{(a-b)})}.$$

Fig. 8.7 shows word polarity scores plotted against occurrence frequency. Using word polarity scores, the top 10 words most indicative of private attorney speech largely include legal terms.

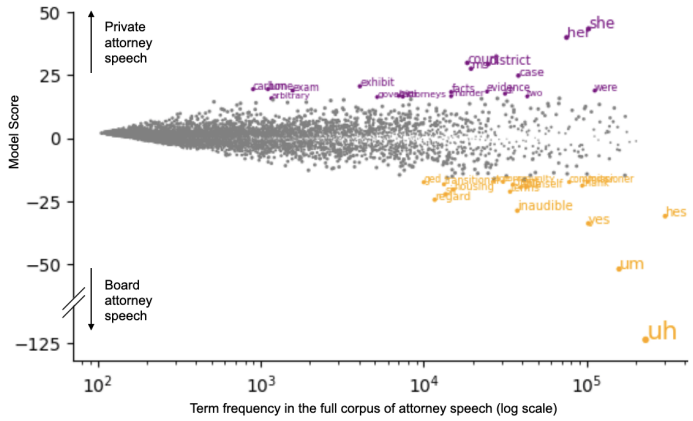

Figure 8.6: Word polarity score for terms most indicative of retained attorneys (in purple) and board-appointed attorneys (in purple), respectively, vs. the occurrence frequency of the word.

Regressions onto the parole hearing outcome

Of the top 10 words, we then assess whether each word significantly predicts hearing outcome in the presence of case factors. Significantly predictive words include *district*, *court* and *exhibit* (all $p < 0.01$ with positive coefficients). Using word polarity scores, the top 10 words indicative of board-appointed attorney speech include *uh, um, yes, sir* and *inaudible*, a term used by hearing transcribers to code unintelligible speech. Of the top 10 words for board-appointed attorneys, we also assess whether each word significantly predicts hearing outcome in the presence of case factors. All ten words significantly predict of the hearing outcome with negative coefficients at the $p < 0.01$ level.

8.3.3 Legal Lexicon

Feature definitions

We designed a custom lexical model to investigate the attorney's usage of specific legal language. We consulted a legal expert who did not have access to the data to devise a list of terms that would reasonably be used by attorneys carrying out their representational duties before the board. The list contains 17 hypothesized terms and covers areas such as evidentiary standards (e.g. *some evidence, reasonable doubt*), procedural terms (e.g. *objection*) and case law establishing precedents for parole in California (e.g. *Lawrence, Shaputis*). For each term, we compute the mean occurrence frequency and the percentage of hearings in which the term is used by the attorney at least once.

Of these terms, 14 of the 17 are used significantly less by board-appointed attorneys in a raw lexical analysis. Mean occurrences differ significantly for all but three of the hypothesized terms (*Miller v Alabama, diminished culpability, reasonable doubt*) at a t-test threshold of $p < 0.05$.

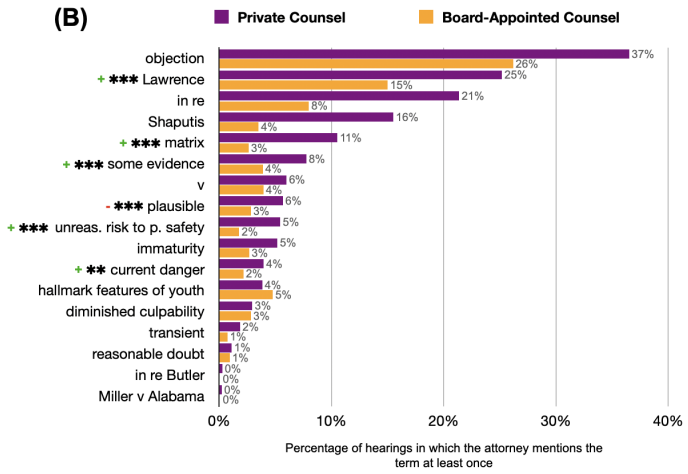

Figure 8.7: Legal term usage by retained (in orange) vs. board-appointed attorneys (in purple). For each term that significantly impacts the parole outcome, it is indicated whether mention of that term increases (+) or decreases (-) the probability of a parole grant in a regression controlling for case factors and speaking time. Wald test significance is marked as *** for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$.

Regressions onto the parole hearing outcome

We run t-tests to check for significance of the differences between the board-appointed and private group. We again check whether each term affects the parole outcome by running independent regressions that assess the impact of the term controlling for case and hearing factors and speaking time. Fig. 8.7 shows that 6 of the 17 terms significantly predict the parole outcome at a $p < 0.01$ level in a logistic regression in the presence of other case factors. A mention of all but one of the 6 terms increases the probability of a parole grant (*plausible* carries a negative coefficient in the regression).

Chapter 9

Conclusion

Parole is the heavy gate at the end of the criminal justice system's long corridor. Its keepers can tip a sentence from fifteen years to fifty. Many of the mechanisms underlying parole have remained opaque not only to the public, but even to governmental oversight bodies. Leveraging the unstructured data recorded in parole hearing transcripts, we employ machine learning tools to extract and analyze parole case factors and shine a light onto this system.

Our analysis identifies several mechanisms that introduce arbitrariness into the parole decision process in California. Factors outside of the candidate's control, such as the historical punitiveness of the commissioner at the time of the hearing and whether the district attorney chooses to attend, have disproportionate weight in explaining the grant outcome after controlling for factors within the candidate's control, such as their educational attainment, participation in rehabilitational programming, disciplinary conduct, and parole plans. There is considerable variability among two key roles in the parole process: commissioners and attorneys. We measure commissioner variability in the presence of non-random assignment to hearings and find significant excess variability in grant rates beyond what should be expected. Using syntactic and lexical measures, we find that board-appointed attorneys speak for less of the hearing and use less legal language, both of which predict denials of parole after controlling for case factors. We uncover that racial disparities limit the voice afforded to non-white parole candidates and their attorneys in the hearing proceedings and that Black and Latinx candidates are significantly less likely to retain a private attorney, which may ultimately halve their chances of being granted parole over candidates who can afford private legal representation. This suggests that race and socioeconomic status may play a significant role in determining parole outcomes in California.

Our work motivates further studies of parole hearing text as data and studies of causal mechanisms in parole. Future studies can replicate our methodology and use advances in NLP to extract additional factors of interest. Both present and future extracted factors can be studied using additional techniques in causal inference to provide a better understanding of the role that race and

social status play in parole decision making. Finally, our work motivates more detailed linguistic studies of the language of parole hearings in a corpus of 5 million pages. A linguistic understanding of parole dynamics could inform many different lines of inquiry.

Our study demonstrates that machine learning can be a powerful tool in bringing reconnaissance to legal decision making. Our approach enables an alternative application for machine learning in criminal law that stands in contrast to its prevailing use as a risk assessment tool. While predictive technology is most commonly applied to analyze the subjects of a decision making process, we propose using it to scrutinize decision making itself Bell et al. [2021]. Our NLP extraction and analysis methodology can be extended to many other legal processes for which only limited structured data is available. Examples include asylum proceedings in an immigration context or social security benefits decisions in an administrative law context.

We believe that two natural next steps follow reconnaissance of parole in California. First, stakeholders in the parole process should consider the implications of our findings for reform proposals in the context of current legal scholarship on parole, case analyses, and the stories of those directly impacted by our justice system. Second, we must identify ways to reconsider the cases of individuals that have been impacted by past inconsistencies and injustices in parole procedures. This dissertation has taken an analytical, historical lens on parole and variability, but parole is a living system. Our data covers the cases of 15,852 individuals, several thousand of whom are still in prison today. The tools we have developed for the present analysis can be used to analyze systems like parole in an ongoing fashion and to identify cases of individuals for review based on specific factors. Our study motivates both comprehensive legislative parole reform and establishing an ongoing review system for parole denials going forward.

Bibliography

R. Abebe, S. Barocas, J. Kleinberg, K. Levy, M. Raghavan, and D. G. Robinson. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 252–260, 2020.

Z. Abedjan, X. Chu, D. Deng, R. C. Fernandez, I. F. Ilyas, M. Ouzzani, P. Papotti, M. Stonebraker, and N. Tang. Detecting data errors: Where are we and what needs to be done? *Proceedings of the VLDB Endowment*, 9(12):993–1004, 2016.

D. S. Abrams, M. Bertrand, and S. Mullainathan. Do judges vary in their treatment of race? *The Journal of Legal Studies*, 41(2):347–383, 2012.

H. Adel, B. Roth, and H. Schütze. Comparing convolutional neural networks to traditional models for slot filling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 828–838, San Diego, California, June 2016. doi: 10.18653/v1/N16-1097.

J. Agley, Y. Xiao, R. Nolan, and L. Golzarri-Arroyo. Quality control questions on Amazon's Mechanical Turk (MTurk): A randomized trial of impact on the USAUDIT, PHQ-9, and GAD-7. *Behavior research methods*, pages 1–13, 2021.

G. Algan and I. Ulusoy. Label noise types and their effects on deep learning. *arXiv:2003.10471*, 2020.

C. Alt, A. Gabryszak, and L. Hennig. TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.142.

K. M. Altenburger and D. E. Ho. Is Yelp actually cleaning up the restaurant industry? A reanalysis on the relative usefulness of consumer reviews. In *The World Wide Web Conference*, pages 2543–2550, 2019.

American Law Institute. Model penal code. *The Institute*, 2019.

E. Amid, M. K. Warmuth, R. Anil, and T. Koren. Robust bi-tempered logistic loss based on Bregman divergences. *Advances in Neural Information Processing Systems*, 32, 2019.

J. M. Anderson, J. R. Kling, and K. Stith. Measuring interjudge sentencing disparity: Before and after the federal sentencing guidelines. *The Journal of Law and Economics*, 42(S1):271–308, 1999.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. In *Ethics of Data and Analytics*, pages 254–264. Auerbach Publications, 2016.

D. Arnold, W. Dobbie, and C. S. Yang. Racial bias in bail decisions. *The Quarterly Journal of Economics*, 133(4):1885–1932, 2018.

D. Arnold, W. Dobbie, and P. Hull. Measuring racial discrimination in algorithms. In *AEA Papers and Proceedings*, volume 111, pages 49–54, 2021.

N. Asher and L. Vieu. Subordinating and coordinating discourse relations. *Lingua*, 115(4):591–610, 2005.

S. Azadi, J. Feng, S. Jegelka, and T. Darrell. Auxiliary image regularization for deep CNNs with noisy labels. *arXiv:1511.07069*, 2015.

D. Baldus. When symbols clash: Reflections on the future of the comparative proportionality review of death sentences. *Seton Hall L. Rev.*, 26:1582, 1995.

D. C. Baldus, G. Woodworth, and C. A. Pulaski. *Equal justice and the death penalty: A legal and empirical analysis*. Upne, 1990.

N. Bar, T. Koren, and R. Giryes. Multiplicative reweighting for robust neural network optimization. *arXiv:2102.12192*, 2021.

C. Barabas, M. Virza, K. Dinakar, J. Ito, and J. Zittrain. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Conference on fairness, accountability and transparency*, pages 62–76. PMLR, 2018.

S. Barocas, M. Hardt, and A. Narayanan. Fairness in machine learning. *NeurIPS tutorial*, 1:2, 2017.

L. Barrett. Reasonably suspicious algorithms: predictive policing at the united states border. *New York University Review of Law and Social Change*, 41:327, 2017.

A. M. Barry-Jester, B. Casselman, and D. Goldstein. The new science of sentencing. *The Marshall Project*, 2015. URL `https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing`.

A. Bäuerle, Á. A. Cabrera, F. Hohman, M. Maher, D. Koski, X. Suau, T. Barik, and D. Moritz. Symphony: Composing interactive interfaces for machine learning. In *CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2022.

A. Beirami, M. Razaviyayn, S. Shahrampour, and V. Tarokh. On optimal generalizability in parametric learning. *Advances in Neural Information Processing Systems*, 30, 2017.

K. Bell. A reparative approach to parole-release decisions. In *Rethinking Punishment in the Era of Mass Incarceration*, pages 162–179. Routledge, 2017.

K. Bell. A stone of hope: Legal and empirical analysis of california juvenile lifer parole decisions. *Harvard Civil Rights-Civil Liberties Law Review*, 54:455, 2019.

K. Bell, J. Hong, N. McKeown, and C. Voss. The recon approach: A new direction for machine learning in criminal law. *Berkeley Technology Law Journal*, 37:102–139, 2021.

I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020.

R. Berk. Accuracy and fairness for juvenile justice risk assessments. *Journal of Empirical Legal Studies*, 16(1):175–194, 2019.

H. Bloch-Wehba. Access to algorithms. *Fordham L. Rev.*, 88:1265, 2019.

J. H. Blume and S. L. Johnson. Unholy parallels between mccleskey v. kemp and plessy v. ferguson: Why mccleskey (still) matters. *Ohio St. J. Crim. L.*, 10:37, 2012.

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv:2108.07258*, 2021.

K. Bowden, J. Wu, S. Oraby, A. Misra, and M. Walker. SlugNERDS: A named entity recognition tool for open domain dialogue systems. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan, May 2018. European Language Resources Association.

S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075.

M. S. Bradley and R. L. Engen. Leaving prison: A multilevel investigation of racial, ethnic, and gender disproportionality in correctional release. *Crime & Delinquency*, 62(2):253–279, 2016.

S. L. Bray. The system of equitable remedies. *UCLA Law Review*, 63:530, 2016.

T. Bynum and R. Paternoster. Discrimination revisited. In *Native Americans, crime, and justice*, pages 228–238. Routledge, 2019.

B. Caldwell. Creating meaningful opportunities for release: Graham, miller and california's youth offender parole hearings. *New York University Review of Law and Social Change*, 40:245, 2016.

California Board of Parole Hearings. Forensic assessment division. URL `http://www.cdcr.ca.gov/BOPH/fad.html,https://perma.cc/2GNW-ST2T`.

M. Chammah. Policing the future. *The Marshall Project*, 2016. URL `https://www.themarshallproject.org/2016/02/03/policing-the-future`.

A. X. Chang and C. Manning. SUTime: A library for recognizing and normalizing time expressions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 3735–3740, Istanbul, Turkey, May 2012. European Language Resources Association.

D. Chen. *Neural Reading Comprehension and Beyond*. PhD thesis, Stanford University, 2018.

D. Chen, A. Fisch, J. Weston, and A. Bordes. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1171.

H. Chen, D. Cai, W. Dai, Z. Dai, and Y. Ding. Charge-based prison term prediction with deep gating network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6362–6367, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1667.

J. Chen, F. Liu, B. Avci, X. Wu, Y. Liang, and S. Jha. Detecting errors and estimating accuracy on unlabeled data with self-training ensembles. *Advances in Neural Information Processing Systems*, 34, 2021a.

P. Chen, J. Ye, G. Chen, J. Zhao, and P.-A. Heng. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11442–11450, 2021b.

T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Y.-H. Chen and J. D. Choi. Character identification on multiparty conversation: Identifying mentions of characters in TV shows. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 90–100, Los Angeles, Sept. 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-3612.

R. Child, S. Gray, A. Radford, and I. Sutskever. Generating long sequences with sparse transformers. *arXiv:1904.10509*, 2019.

J. D. Choi and H. Y. Chen. SemEval 2018 task 4: Character identification on multiparty dialogues. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 57–64, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1007.

C. Clark and M. Gardner. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1078.

C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300.

T. H. Cohen. Who is better at defending criminals? does type of defense attorney matter in terms of producing favorable case outcomes. *Criminal Justice Policy Review*, 25(1):29–58, 2014.

Committee on Revision of the Penal Code. Memorandum 2020-15 of the committee on revision of the penal code. Technical report, 2020. URL `http://www.clrc.ca.gov/CRPC/Pub/Memos/CRPC20-15.pdf`.

S. Corbett-Davies and S. Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv:1808.00023*, 2018.

S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 797–806, 2017.

Z. Dai, Z. Li, and L. Han. Bonebert: A bert-based automated information extraction system of radiology reports for bone fracture detection and diagnosis. In *IDA*, pages 263–274, 2021.

C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, and J. Kleinberg. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708, 2012.

R. Das, S. Dhuliawala, M. Zaheer, and A. McCallum. Multi-step retriever-reader interaction for scalable open-domain question answering. In *International Conference on Learning Representations*, 2019.

K. C. Davis. *Discretionary justice: A preliminary inquiry*. LSU Press, 1969.

G. Dawson and R. Polikar. Rethinking noisy label models: Labeler-dependent noise with adversarial awareness. *arXiv preprint arXiv:2105.14083*, 2021.

DCAI Workshop. NeurIPS data-centric AI workshop. *Advances in Neural Information Processing Systems 2021 Data-Centric AI Workshop*, 2021. URL `https://datacentricai.org/neurips21/`.

A. De Raadt, M. J. Warrens, R. J. Bosker, and H. A. Kiers. Kappa coefficients for missing data. *Educational and psychological measurement*, 79(3):558–576, 2019.

M. Debruyne, M. Hubert, and J. A. Suykens. Model selection in kernel based regression using the influence function. *Journal of machine learning research.-Cambridge, Mass.*, 9:2377–2400, 2008.

A. Deeks. The judicial demand for explainable artificial intelligence. *Columbia Law Review*, 119(7): 1829–1850, 2019.

S. J. Delany, N. Segata, and B. Mac Namee. Profiling instances in noise reduction. *Knowledge-Based Systems*, 31:28–40, 2012.

S. A. Dennis, B. M. Goodson, and C. A. Pearson. Online worker fraud and evolving threats to the integrity of MTurk data: A discussion of virtual private servers and the limitations of IP-based screening procedures. *Behavioral Research in Accounting*, 32(1):119–134, 2020.

S. Desai and G. Durrett. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.21. URL `https://aclanthology.org/2020.emnlp-main.21`.

S. L. Desmarais and S. A. Zottola. Violence risk assessment: Current status and contemporary issues. *Marq. L. Rev.*, 103:793, 2019.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://www.aclweb.org/anthology/N19-1423`.

C. D'ignazio and L. F. Klein. *Data feminism*. MIT press, 2020.

J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*, 2019.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

J. M. Eaglin. Constructing recidivism risk. *Emory LJ*, 67:59, 2017.

T. Eisenberg. The origins, nature, and promise of empirical legal studies and a response to concerns. *U. Ill. L. Rev.*, page 1713, 2011.

J. K. Elek, R. K. Warren, and P. M. Casey. *Using risk and needs assessment information at sentencing: Observations from ten jurisdictions*. National Center for State Courts, 2015.

D. F. Engstrom and D. E. Ho. Algorithmic accountability in the administrative state. *Yale J. on Reg.*, 37:800, 2020.

D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*, pages 160–171. PMLR, 2018.

H. Fair and R. Walmsley. *World prison population list*. Institute for Crime & Justice Policy Research, 2021. URL `https://www.prisonstudies.org/sites/default/files/resources/downloads/world_prison_population_list_13th_edition.pdf`.

H. Fang, H. Cheng, M. Sap, E. Clark, A. Holtzman, Y. Choi, N. A. Smith, and M. Ostendorf. Sounding board: A user-centric and content-driven social chatbot. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 96–100, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-5020. URL `https://www.aclweb.org/anthology/N18-5020`.

A. G. Ferguson. Illuminating black data policing. *Ohio St. J. Crim. L.*, 15:503, 2017.

S. E. Finch, J. D. Finch, A. Ahmadvand, I. Choi, X. Dong, R. Qi, H. Sahijwani, S. Volokhin, Z. Wang, Z. Wang, and J. D. Choi. Emora: An inquisitive social chatbot who cares for you. In *3rd Proceedings of Alexa Prize*, 2020.

B. Frenay and M. Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014. doi: 10.1109/TNNLS.2013.2292894.

S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 329–338, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287589. URL https://doi.org/10.1145/3287560.3287589.

D. R. Friedman and J. M. Robinson. Rebutting the presumption: an empirical analysis of parole deferrals under marsy's law. *Stan. L. Rev.*, 66:173, 2014.

R. G. Fryer Jr. An empirical analysis of racial differences in police use of force. *Journal of Political Economy*, 127(3):1210–1261, 2019.

Furman vs. Georgia. 408 U.S. 238, 365. *U.S.*, 1972.

D. Gamberger, N. Lavrac, and S. Dzeroski. Noise detection and elimination in data preprocessing: experiments in medical domains. *Applied artificial intelligence*, 14(2):205–223, 2000.

S. Geisser. The predictive sample reuse method with applications. *Journal of the American statistical Association*, 70(350):320–328, 1975.

A. Gelman, J. Fagan, and A. Kiss. An analysis of the new york city police department's "stop-and-frisk" policy in the context of claims of racial bias. *Journal of the American statistical association*, 102(479):813–823, 2007.

C. S. Germain. *The Doctor and Student*. 1518. URL https://lonang.com/wp-content/download/DoctorAndStudent.pdf.

N. Ghandnoosh. US prison decline: Insufficient to undo mass incarceration. *The Sentencing Project*, 2020. URL http://arks.princeton.edu/ark:/88435/dsp01rn3014452.

T. Gillespie. Algorithm [draft] [#digitalkeywords]. Online, June 2014. ACulture Digitally. URL https://culturedigitally.org/2014/06/algorithm-draft-digitalkeyword/.

K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-speech tagging for Twitter: Annotation, features, and experiments. Technical report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science, 2010.

R. Giordano, W. Stephenson, R. Liu, M. Jordan, and T. Broderick. A swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147. PMLR, 2019.

S. Goel, J. M. Rao, and R. Shroff. Personalized risk assessments in the criminal justice system. *American Economic Review*, 106(5):119–23, 2016.

L. E. Goodman. In defense of federal judicial sentencing. *Calif. L. Rev.*, 46:497, 1958.

K. Greenawalt. Discretion and judicial decision: The elusive quest for the fetters that bind judges. *Colum. L. Rev.*, 75:359, 1975.

J. Greene and I. Dalke. "you're still an angry man": Parole boards and logics of criminalized masculinity. *Theoretical Criminology*, page 1362480620910222, 2020.

J. Grimmer and B. M. Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297, 2013.

A. Gross. History, race, and prediction: Comments on harcourt's against prediction. *Law & Social Inquiry*, 33(1):233–242, 2008.

S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL `https://aclanthology.org/2020.acl-main.740`.

P. Gustafson. *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. CRC Press, 2003.

B. E. Harcourt. Against prediction: Sentencing, policing, and punishing in an actuarial age. *Chicago Public Law and Legal Theory Working Paper*, (94), 2005.

H. L. Hart. Discretion. *Harvard Law Review*, 127(2):652–665, 2013.

R. D. Hartley, H. V. Miller, and C. Spohn. Do you get what you pay for? type of counsel and its effect on criminal court outcomes. *Journal of Criminal Justice*, 38(5):1063–1070, 2010.

P. He, J. Gao, and W. Chen. DeBERTaV3: Improving DeBERTa using Electra-style pre-training with gradient-disentangled embedding sharing. *arXiv:2111.09543*, 2021.

A. M. Heinz, J. P. Heinz, S. J. Senderowitz, and M. A. Vance. Sentencing by parole board: An evaluation. *The Journal of Criminal Law and Criminology (1973-)*, 67(1):1–31, 1976.

D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017.

D. Hendrycks, M. Mazeika, D. Wilson, and K. Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. *Advances in neural information processing systems*, 31, 2018.

D. Hendrycks, X. Liu, E. Wallace, A. Dziedzic, R. Krishnan, and D. Song. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.244. URL `https://aclanthology.org/2020.acl-main.244`.

D. Hendrycks, C. Burns, A. Chen, and S. Ball. CUAD: An expert-annotated nlp dataset for legal contract review. *Advances in Neural Information Processing Systems*, 2021.

K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, 2015.

F. Hill, A. Bordes, S. Chopra, and J. Weston. The goldilocks principle: Reading children's books with explicit memory representations. In Y. Bengio and Y. LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL `http://arxiv.org/abs/1511.02301`.

D. E. Ho. Does peer review work? an experiment of experimentalism. *Stan. L. Rev.*, 69:1, 2017.

D. E. Ho and A. Xiang. Affirmative algorithms: The legal grounds for fairness as awareness. *University of Chicago Law Review Online*, 2020.

T. K. Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

J. M. Hofman, D. J. Watts, S. Athey, F. Garip, T. L. Griffiths, J. Kleinberg, H. Margetts, S. Mullainathan, M. J. Salganik, S. Vazire, et al. Integrating explanation and prediction in computational social science. *Nature*, 595:1–8, 2021.

J. Hong, D. Chong, and C. Manning. Learning from limited labels for long legal dialogue. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 190–204, Punta Cana, Dominican Republic, Nov. 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.nllp-1.20. URL `https://aclanthology.org/2021.nllp-1.20`.

J. Hong, C. Voss, and C. Manning. Challenges for information extraction from dialogue in criminal law. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 71–81, Online,

Aug. 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.nlp4posimpact-1.8. URL https://aclanthology.org/2021.nlp4posimpact-1.8.

D. Hovy, B. Plank, and A. Søgaard. Experiments with crowdsourced re-annotation of a POS tagging data set. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 377–382, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-2062. URL https://aclanthology.org/P14-2062.

Z. Hu, X. Li, C. Tu, Z. Liu, and M. Sun. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 487–498, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1041.

J. Huang, L. Qu, R. Jia, and B. Zhao. O2u-net: A simple noisy label detection approach for deep neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3326–3334, 2019.

B. M. Huebner and T. S. Bynum. The role of race and ethnicity in parole decisions. *Criminology*, 46(4):907–938, 2008.

A. Z. Huq. A right to a human decision. *Va. L. Rev.*, 106:611, 2020.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

W. S. Isaac. Hope, hype, and fear: the promise and potential pitfalls of artificial intelligence in criminal justice. *Ohio St. J. Crim. L.*, 15:543, 2017.

J. Isard. Under the cloak of brain science: Risk assessments, parole, and the powerful guise of objectivity. *Calif. L. Rev.*, 105:1223, 2017.

J. B. Jacobs. The uneasy truce between law and equity in modern business enterprise jurisprudence. *Del. L. Rev.*, 8:1–9, 2005.

L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pages 2304–2313. PMLR, 2018.

Y. Jiang, V. Nagarajan, C. Baek, and J. Z. Kolter. Assessing generalization of SGD via disagreement. In *International Conference on Machine Learning*, 2022.

E. E. Joh. Feeding the machine: Policing, crime data, & algorithms. *Wm. & Mary Bill Rts. J.*, 26:287, 2017.

A. Joulin, L. v. d. Maaten, A. Jabri, and N. Vasilache. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pages 67–84. Springer, 2016.

A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics, April 2017.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang. A comparative study on transformer vs RNN in speech applications. pages 449–456. IEEE, 2019.

V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL `https://aclanthology.org/2020.emnlp-main.550`.

S. Karthik, J. Revaud, and B. Chidlovskii. Learning from long-tailed data with noisy labels. *arXiv:2108.11096*, 2021.

R. Kennedy, S. Clifford, T. Burleigh, P. D. Waggoner, R. Jewell, and N. J. Winter. The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods*, 8(4):614–629, 2020.

T. Kim, J. Ko, J. Choi, S.-Y. Yun, et al. FINE samples for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34, 2021.

N. Kitaev, Ł. Kaiser, and A. Levskaya. Reformer: The efficient transformer. *arXiv:2001.04451*, 2020.

S. Klein, J. Petersilia, and S. Turner. Race and imprisonment decisions in california. *Science*, 247 (4944):812–816, 1990.

T. Klein and M. Nabi. Learning to answer by learning to ask: Getting the best of gpt-2 and bert worlds. *arXiv:1911.02365*, 2019.

J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2018a.

J. Kleinberg, J. Ludwig, S. Mullainathan, and A. Rambachan. Algorithmic fairness. In *AEA papers and proceedings*, volume 108, pages 22–27, 2018b.

J. Kleinberg, J. Ludwig, S. Mullainathan, and C. R. Sunstein. Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10:113–174, 2018c.

J. Kreutzer, I. Caswell, L. Wang, A. Wahab, D. van Esch, N. Ulzii-Orshikh, A. Tapo, N. Subramani, A. Sokolov, C. Sikasote, et al. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022.

K. Krishna, A. Roy, and M. Iyyer. Hurdles to progress in long-form question answering. *arXiv:2103.06332*, 2021.

A. Kumar and E. Amid. Constrained instance and class reweighting for robust learning under label noise. *arXiv:2111.05428*, 2021.

V. Kuperman, H. Stadthagen-Gonzalez, and M. Brysbaert. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44(4):978–990, 2012.

T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1082. URL `https://aclanthology.org/D17-1082`.

D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. Computational social science. *Science*, 323(5915):721, 2009.

G. Lee, S.-w. Hwang, and H. Cho. Squad2-cr: Semi-supervised annotation for cause and rationales for unanswerability in squad 2.0. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5425–5432, 2020.

D. Lehr and P. Ohm. Playing with the data: what legal scholars should learn about machine learning. *UCDL Rev.*, 51:653, 2017.

K. Leins, J. H. Lau, and T. Baldwin. Give me convenience and give her death: Who should decide what uses of NLP are appropriate, and on what basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913. Association for Computational Linguistics, July 2020. doi: 10.18653/v1/2020.acl-main.261. URL `https://www.aclweb.org/anthology/2020.acl-main.261`.

O. Levy, M. Seo, E. Choi, and L. Zettlemoyer. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/K17-1034. URL `https://www.aclweb.org/anthology/K17-1034`.

K. Liang, A. Chau, Y. Li, X. Lu, D. Yu, M. Zhou, I. Jain, S. Davidson, J. Arnold, M. Nguyen, and Z. Yu. Gunrock 2.0: A user adaptive social conversational system. In *3rd Proceedings of Alexa Prize*, 2020.

E. Lieberman, J.-B. Michel, J. Jackson, T. Tang, and M. A. Nowak. Quantifying the evolutionary dynamics of language. *Nature*, 449(7163):713–716, 2007.

A. Y. Ling, A. W. Kurian, J. L. Caswell-Jin, G. W. Sledge Jr, N. H. Shah, and S. R. Tamang. A semi-supervised machine learning approach to detecting recurrent metastatic breast cancer cases using linked cancer registry and electronic medical record data. *arXiv preprint arXiv:1901.05958*, 2019.

Y. Liu and H. Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International Conference on Machine Learning*, pages 6226–6236. PMLR, 2020.

Y. Liu, S. Jiang, and S. Liao. Efficient approximation of cross-validation for kernel methods using bouligand influence function. In *International Conference on Machine Learning*, pages 324–332. PMLR, 2014.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Lockett vs. Ohio. 438 U.S. 586, 606. *U.S.*, 1978.

F. Luo, A. Nagesh, R. Sharp, and M. Surdeanu. Semi-supervised teacher-student architecture for relation extraction. In *Proceedings of the Third Workshop on Structured Prediction for NLP*, pages 29–37, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1505. URL `https://www.aclweb.org/anthology/W19-1505`.

X. Ma, H. Huang, Y. Wang, S. Romano, S. Erfani, and J. Bailey. Normalized loss functions for deep learning with noisy labels. In *International Conference on Machine Learning*, pages 6543–6553. PMLR, 2020.

A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `https://aclanthology.org/P11-1015`.

A. Maheshwari, O. Chatterjee, K. Killamsetty, R. Iyer, and G. Ramakrishnan. Data programming using semi-supervision and subset selection. *arXiv preprint arXiv:2008.09887*, 2020.

H. H. Malik and V. S. Bhardwaj. Automatic training data cleaning for text classification. In *2011 IEEE 11th international conference on data mining workshops*, pages 442–449. IEEE, 2011.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993. URL `https://aclanthology.org/J93-2004`.

J. L. Mashaw. Administrative due process: The quest for a dignitary theory. *Bul Rev.*, 61:885, 1981.

S. G. Mayson. Bias in, bias out. *Yale Law Journal*, 128:2218, 2018.

J. McAuley, C. Targett, Q. Shi, and A. van den Hengel. Image-based recommendations on styles and substitutes. In *SIGIR 2015*, page 43–52, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336215. doi: 10.1145/2766462.2767755. URL `https://doi.org/10.1145/2766462.2767755`.

McCleskey vs. Kemp. 481 U.S. 279, 87. *U.S.*, 1987.

S. Mechoulan and N. Sahuguet. Assessing racial disparities in parole release. *The Journal of Legal Studies*, 44(1):39–74, 2015.

S. Mehta. False hope: How parole systems fail youth serving extreme sentences. Technical report, 2016. URL `https://perma.cc/5CF3-BC3W`.

A. M. Mellis and W. K. Bickel. Mechanical Turk data collection in addiction research: Utility, concerns and best practices. *Addiction*, 115(10):1960–1968, 2020.

C. Metz and A. Satariano. An algorithm that grants freedom, or takes it away. *New York Times*, 2020.

J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, et al. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

T. More. *Utopia*. 1551.

A. Moss and L. Litman. After the bot scare: Understanding what's been happening with data collection on MTurk and how to stop it. *CloudResearch*, 2018. URL `https://www.cloudresearch.com/resources/blog/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-`

N. M. Müller and K. Markert. Identifying mislabeled instances in classification datasets. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

A. Nellis. No end in sight: America's enduring reliance on life imprisonment. *The Sentencing Project*, 2021. URL `https://www.sentencingproject.org/app/uploads/2022/08/No-End-in-Sight-Americas-Enduring-Reliance-on-Life-Imprisonment.pdf`.

T. H. Nguyen and R. Grishman. Combining neural networks and log-linear models to improve relation extraction. arXiv:1511.05926, 2015.

C. Northcutt, L. Jiang, and I. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021a.

C. G. Northcutt, A. Athalye, and J. Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021b.

M. C. Nussbaum. Equity and mercy. *Philosophy & Public Affairs*, pages 83–125, 1993.

B. Oudekerk and D. Kaeble. Probation and parole in the united states, 2019. *Washington DC: US Department of Justice*, 2021.

B. M. Ovalle. Examining the quality of representation by public defenders compared to private attorneys. *The Mid-Southern Journal of Criminal Justice*, 2(1):2, 2021.

O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `https://aclanthology.org/N13-1039`.

F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller. Language models as knowledge bases? 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)*, pages 2463–2473, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL `https://aclanthology.org/D19-1250`.

E. Pierson, C. Simoiu, J. Overgoor, S. Corbett-Davies, D. Jenson, A. Shoemaker, V. Ramachandran, P. Barghouty, C. Phillips, R. Shroff, et al. A large-scale analysis of racial disparities in police stops across the united states. *Nature human behaviour*, 4(7):736–745, 2020.

B. Plank, D. Hovy, and A. Søgaard. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, 2014.

E. D. Poole and R. M. Regoli. Race, institutional rule breaking, and disciplinary response: A study of discretionary decision making in prison. *Law & Soc'y Rev.*, 14:931, 1979.

R. A. Posner and A. H. Yoon. What judges think of the quality of legal representation. *Stan. L. Rev.*, 63:317, 2010.

G. J. Postema. Dilemmas of Discretion: Equity and Mercy. In *Law's Rule: The Nature, Value, and Viability of the Rule of Law*. Oxford University Press, 01 2023. ISBN 9780190645342. doi: 10.1093/oso/9780190645342.003.0011. URL `https://doi.org/10.1093/oso/9780190645342.003.0011`.

J. J. Rachlinski, S. L. Johnson, A. J. Wistrich, and C. Guthrie. Does unconscious racial bias affect trial judges. *Notre Dame L. Rev.*, 84:1195, 2008.

K. R. Rad and A. Maleki. A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):965–996, 2020.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. OpenAI, 2019. URL `https://openai.com/blog/better-language-models/`.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, Nov. 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL `https://www.aclweb.org/anthology/D16-1264`.

A. Ratner, C. De Sa, S. Wu, D. Selsam, and C. Ré. Data programming: Creating large training sets, quickly. *Advances in Neural Information Processing Systems*, 29:3567, 2016.

A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, 11(3):269–282, 2017.

A. Ratner, B. Hancock, J. Dunnmon, R. Goldman, and C. Ré. Snorkel metal: Weak supervision for multi-task learning. In *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*, pages 1–4, 2018.

R. M. Re and A. Solow-Niederman. Developing artificially intelligent justice. *Stan. Tech. L. Rev.*, 22:242, 2019.

S. Reddy, D. Chen, and C. D. Manning. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

T. C. Redman. The impact of poor data quality on the typical enterprise. *Communications of the ACM*, 41(2):79–82, 1998.

S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

P. D. Reingold and K. A. Thomas. From grace to grids: Rethinking due process protections for parole. *J. Crim. L. & Criminology*, 107(2):213–292, 2017.

F. Reiss, H. Xu, B. Cutler, K. Muthuraman, and Z. Eichenberger. Identifying incorrect labels in the conll-2003 corpus. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 215–226, 2020.

K. R. Reitz and E. E. Rhine. Parole release and supervision: critical drivers of American prison policy. *Annual Review of Criminology*, 3:281–298, 2020.

J. Renaud. Grading the parole release systems of all 50 states. *Prison Policy Initiative*, 2019.

E. E. Rhine, J. Petersilia, and K. R. Reitz. The future of parole release. *Crime and Justice*, 46(1): 279–338, 2017.

M. Richardson, C. J. Burges, and E. Renshaw. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA, Oct. 2013. Association for Computational Linguistics. URL `https://aclanthology.org/D13-1020`.

K. T. Rodolfa, E. Salomon, L. Haynes, I. H. Mendieta, J. Larson, and R. Ghani. Case study: Predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 142–153, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372863. URL `https://doi.org/10.1145/3351095.3372863`.

D. Rolnick, A. Veit, S. Belongie, and N. Shavit. Deep learning is robust to massive label noise. *arXiv:1705.10694*, 2017.

A. Roy, M. Saffar, A. Vaswani, and D. Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68, 2021. doi: 10.1162/tacl_a_00353. URL `https://aclanthology.org/2021.tacl-1.4`.

D. L. Rubinfeld. Reference guide on multiple regression. *Reference manual on scientific evidence*, 179:425–469, 2000.

T. Saito and M. Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.

J. Samuel, G. Rozzi, and R. Palle. The dark side of sentiment analysis: An exploratory review using lexicons, dictionaries, and a statistical monkey and chimp. *Dictionaries, and a Statistical Monkey and Chimp.(January 6, 2022)*, 2022.

J. Sánchez-Monedero, L. Dencik, and L. Edwards. What does it mean to 'solve' the problem of discrimination in hiring? social, technical and legal perspectives from the uk on automated hiring systems. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 458–468, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372849. URL `https://doi.org/10.1145/3351095.3372849`.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv:1910.01108, 2019. version 4.

A. Saravanos, S. Zervoudakis, D. Zheng, N. Stott, B. Hawryluk, and D. Delfino. The hidden cost of using Amazon Mechanical Turk for research. In *International Conference on Human-Computer Interaction*, pages 147–164. Springer, 2021.

M. Schuster and K. Nakajima. Japanese and korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE, 2012.

R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. *arXiv:1508.07909*, 2015.

A. Shapiro. Reform predictive policing. *Nature*, 541(7638):458–460, 2017.

H. J. Singer and K. W. Caves. Applied econometrics: When can an omitted variable invalidate a regression? *The AntiTrust Source*, 17(3):9, 2017.

D. Slater. How to get out of prison: Can you talk your way out of a life sentence. *New York Times Magazine*, 1(1):32, 2020.

B. Sluban, D. Gamberger, and N. Lavrač. Ensemble-based noise detection: noise ranking and visual performance evaluation. *Data mining and knowledge discovery*, 28(2):265–303, 2014.

A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020.

H. E. Smith. Why fiduciary law is equitable. *Philosophical Foundations of Fiduciary Law, Oxford University Press, Forthcoming, Harvard Public Law Working Paper*, (13-36), 2013.

R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii, Oct. 2008. Association for Computational Linguistics. URL `https://aclanthology.org/D08-1027`.

L. B. Solum. Procedural justice. *Southern California Law Review*, 78:181, 2004.

H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

P. Stancil. Substantive equality and procedural justice. *Iowa L. Rev.*, 102:1633, 2016.

S. B. Starr. Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford law review*, pages 803–872, 2014.

State vs. Loomis. 881 N.W.2d 749, 755. *Wisconsin*, 2016.

M. Stevenson. Assessing risk assessment in action. *Minn. L. Rev.*, 103:303, 2018.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133, 1974.

K. J. Strandburg. Rulemaking and inscrutable automated decision tools. *Columbia Law Review*, 119(7):1851–1886, 2019.

C. Sun, X. Qiu, Y. Xu, and X. Huang. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer, 2019a.

K. Sun, D. Yu, J. Chen, D. Yu, Y. Choi, and C. Cardie. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231, 2019b.

Superior Court of California in and for the County of San Francisco. Brodheim v. California Department of Corrections and Rehabilitation, 2020a. URL `https://www.eff.org/document/order-brodheim-v-cdcr-voss-v-cdcr-companion-case`.

Superior Court of California in and for the County of San Francisco. Brodheim v. California Department of Corrections and Rehabilitation, 2020b. URL `https://www.eff.org/document/order-voss-v-cdcr`.

M. Surdeanu, R. Nallapati, G. Gregory, J. Walker, and C. D. Manning. Risk analysis for intellectual property litigation. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law*, pages 116–120, 2011.

R. Swanson, B. Ecker, and M. Walker. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226, Prague, Czech Republic, Sept. 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-4631. URL https://www.aclweb.org/anthology/W15-4631.

B. Szmrecsanyi. On operationalizing syntactic complexity. In *Le poids des mots. Proceedings of the 7th international conference on textual data statistical analysis. Louvain-la-Neuve*, volume 2, pages 1032–1039, 2004.

J. Tasioulas. The paradox of equity. *The Cambridge Law Journal*, 55(3):456–469, 1996.

K. Thomas and P. Reingold. From grace to grids. *The Journal of Criminal Law and Criminology (1973-)*, 107(2):213–252, 2017.

J. Thongkam, G. Xu, Y. Zhang, and F. Huang. Support vector machine for outlier detection in breast cancer survivability prediction. In *Asia-Pacific Web Conference*, pages 99–109. Springer, 2008.

C. Tibbitts. Success or failure on parole can be predicted: A study of the records of 3,000 youths paroled from the illinois state reformatory. *Am. Inst. Crim. L. & Criminology*, 22:11, 1931.

G. Todd, C. Voss, and J. Hong. Unsupervised anomaly detection in parole hearings using language models. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 66–71, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpcss-1.8. URL https://aclanthology.org/2020.nlpcss-1.8.

S. Tonn. Can AI help judges make the bail system fairer and safer? *Stanford School of Engineering Magazine*, 2019. URL https://engineering.stanford.edu/magazine/article/can-ai-help-judges-make-bail-system-fairer-and-safer.

M. Tonry. Predictions of dangerousness in sentencing: Déjà vu all over again. *Crime and Justice*, 48(1):439–482, 2019.

T. R. Tyler. Why people obey the law: Procedural justice. *Legitimacy, and Compliance*, 1990.

T. R. Tyler. Procedural justice, legitimacy, and the effective rule of law. *Crime and justice*, 30: 283–357, 2003.

S. S. Ulmer. Quantitative analysis of judicial processes: Some practical and theoretical applications. *Law and Contemporary Problems*, 28(1):164–184, 1963.

G. Van Eijk. Socioeconomic marginality in sentencing: The built-in bias in risk assessment tools and the reproduction of social inequality. *Punishment & Society*, 19(4):463–481, 2017.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

R. Voigt, N. P. Camp, V. Prabhakaran, W. L. Hamilton, R. C. Hetey, C. M. Griffiths, D. Jurgens, D. Jurafsky, and J. L. Eberhardt. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25):6521–6526, 2017.

R. Walmsley. Global incarceration and prison trends. In *Forum on Crime and Society*, volume 3, pages 65–78, 2003.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL `https://aclanthology.org/W18-5446`.

P. Wang, Z. Deng, and R. Cui. Tdjee: A document-level joint model for financial event extraction. *Electronics*, 10(7):824, 2021.

M. Watson, B. A. S. Hasan, and N. Al Moubayed. Agree to disagree: When deep learning models with identical architectures produce distinct explanations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 875–884, 2022.

D. J. Watts. Should social science be more solution-oriented? *Nature Human Behaviour*, 1(1):1–5, 2017.

J. Wei, Z. Zhu, H. Cheng, T. Liu, G. Niu, and Y. Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=TBWA6PLJZQm`.

R. Weisberg, D. A. Mukamal, and J. D. Segall. *Life in limbo: An examination of parole release for prisoners serving life sentences with the possibility of parole in California*. Stanford Law School, Stanford Criminal Justice Center Stanford, CA, 2011. URL `https://perma.cc/3WUL-SGCB`.

R. Wexler. Life, liberty, and trade secrets: Intellectual property in the criminal justice system. *Stanford Law Review*, 70, 2017.

V. Wheway. Using boosting to detect noisy data. In *Pacific Rim International Conference on Artificial Intelligence*, pages 123–130. Springer, 2000.

E. Widra and T. Herring. *States of Incarceration: The Global Context 2021*. Prison Policy Initiative, 2021. URL `https://www.prisonpolicy.org/global/2021.html`.

M. R. Williams. The effectiveness of public defenders in four florida counties. *Journal of Criminal Justice*, 41(4):205–212, 2013.

A. Wilson, M. Kasy, and L. Mackey. Approximate cross-validation: Guarantees for model assessment and selection. In *International Conference on Artificial Intelligence and Statistics*, pages 4530–4540. PMLR, 2020.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`.

L. Xu, J. Liu, X. Pan, X. Lu, and X. Hou. DataCLUE: A benchmark suite for data-centric NLP. *arXiv:2111.08647*, 2021a.

R. Xu, T. Liu, L. Li, and B. Chang. Document-level event extraction via heterogeneous graph-based interaction model with a tracker. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3533–3546, Online, Aug. 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.274. URL `https://aclanthology.org/2021.acl-long.274`.

Z. Yang and J. D. Choi. FriendsQA: Open-domain question answering on TV show transcripts. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 188–197, Stockholm, Sweden, Sept. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5923. URL `https://www.aclweb.org/anthology/W19-5923`.

Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL `https://aclanthology.org/D18-1259`.

Y. Yao, D. Ye, P. Li, X. Han, Y. Lin, Z. Liu, Z. Liu, L. Huang, J. Zhou, and M. Sun. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, 2019.

K. M. Young. Parole hearings and victims' rights: Implementation, ambiguity, and reform. *Conn. L. Rev.*, 49:431, 2016.

K. M. Young, D. A. Mukamal, and T. Favre-Bulle. Predicting parole grants: An analysis of suitability hearings for california's lifer inmates. *Fed. Sent'g Rep.*, 28:268, 2015.

D. Yu, K. Sun, C. Cardie, and D. Yu. Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4927–4940. Association for Computational Linguistics, July 2020. doi: 10.18653/v1/2020.acl-main.444. URL `https://www.aclweb.org/anthology/2020.acl-main.444`.

M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed. Big bird: Transformers for longer sequences. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf`.

J. Zhang, V. S. Sheng, Q. Li, J. Wu, and X. Wu. Consensus algorithms for biased labeling in crowdsourcing. *Information Sciences*, 382:254–273, 2017.

J. Zhang, Y. Zhao, M. Saleh, and P. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR, 2020.

S. Zhang, L. He, E. Dragut, and S. Vucetic. How to invest my time: Lessons from human-in-the-loop entity extraction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2305–2313, 2019.

T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SkeHuCVFDr`.

W. Zhang and K. Stratos. Understanding hard negatives in noise contrastive estimation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1090–1101, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.86. URL `https://aclanthology.org/2021.naacl-main.86`.

S. Zheng, W. Cao, W. Xu, and J. Bian. Doc2EDAG: An end-to-end document-level framework for Chinese financial event extraction. 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), pages 337–346, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1032. URL https://aclanthology.org/D19-1032.

H. Zhong, Z. Guo, C. Tu, C. Xiao, Z. Liu, and M. Sun. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1390. URL https://www.aclweb.org/anthology/D18-1390.

W. Zhou and M. Chen. Learning from noisy labels for entity-centric information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5381–5392, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.437. URL https://aclanthology.org/2021.emnlp-main.437.

D. Zhu, M. A. Hedderich, F. Zhai, D. Adelani, and D. Klakow. Is BERT robust to label noise? A study on learning with noisy labels in text classification. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 62–67, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.insights-1.8. URL https://aclanthology.org/2022.insights-1.8.