THE RECON APPROACH: A NEW DIRECTION FOR MACHINE LEARNING IN CRIMINAL LAW

Kristen Bell[†], *Jenny Hong*[††], *Nick McKeown*[†††] *& Catalin Voss*[‡]

ABSTRACT

Most applications of machine learning in criminal law focus on making predictions about people and using those predictions to guide decisions. For example, judges use risk assessment tools to predict the likelihood of future violence when making decisions about whom to detain pre-trial. Whereas this predictive technology analyzes people about whom decisions are made, we propose a new direction for machine learning that scrutinizes decision-making itself. Our aim is not to predict behavior but to provide the public with data-driven opportunities to improve the fairness and consistency of human discretionary judgment. We call our approach the Recon Approach because it encompasses two functions: reconnaissance and reconsideration. Reconnaissance harnesses natural language processing to cull through thousands of hearing transcripts and illuminate factors that appear to have influenced decisions at those hearings. Reconsideration uses modeling techniques to identify cases that appear anomalous in a way that warrants a closer review of those decisions. Reconnaissance reveals patterns that may show systemic problems across a set of decisions; reconsideration flags potential errors or injustices in individual cases. As a team of computer scientists and legal scholars, we describe our early work to apply the Recon Approach to parole-release decisions in California. Drawing on that work, we discuss challenges to the Recon Approach as well as its potential to apply to sentencing and other discretionary decision-making contexts within and beyond criminal law.

DOI: https://doi.org/10.15779/Z38NC5SD2Z

† Kristen Bell is Assistant Professor at University of Oregon School of Law.

†† Jenny Hong is a PhD student at Stanford University.

††† Nick McKeown is Professor of Computer Science and Electrical Engineering at Stanford University.

‡ Catalin Voss is a PhD student at Stanford University. The authors would like to thank the following for their advice and feedback in early development of this work: Robert Weisberg, Terry Winograd, and Daniel Ho. The authors would also like to thank Stanford HAI and the Google PhD Fellowship Program for funding and the Electronic Frontier Foundation for legal assistance in helping them develop Project Recon.

TABLE OF CONTENTS

I. INTRODUCTION

Computer scientists are increasingly engaged in developing machine-learning technology for criminal law. Much of that technology is designed to predict the likelihood that an individual will commit violence in the future. The intended users of this predictive technology include police officers deciding

whom to stop,[1] judges deciding whom to retain in custody pre-trial,[2] judges deciding what sentence to impose,[3] and parole boards deciding whom to keep imprisoned.[4] This type of technological development follows what we call the Predictive Approach. This approach tends to channel technological development narrowly because it is designed to assess those who are processed through the legal system, although it generally neglects to assess those who are making the decisions. Working together as computer scientists and legal scholars, we propose an alternative and additional path for machine learning that shifts the focus from the people about whom decisions are made to the decision-making itself. We call this path the Recon Approach.

The Recon Approach 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.[5] 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

1. *See, e.g.*, Andrew Guthrie Ferguson, *Illuminating Black Data Policing*, 15 OHIO ST. J. CRIM. L. 503, 505 (2018) (describing predictive policing technologies); Lindsey Barrett, *Reasonably Suspicious Algorithms: Predictive Policing at the United States Border*, 41 N.Y.U. REV. L. & SOC. CHANGE 327, 335 (2017); Sharad Goel, Justin M. Rao & Ravi Shroff, *Personalized Risk Assessments in the Criminal Justice System*, 106 AM. ECON. REV.: PAPERS & PROC. 119 (2016).

2. *See, e.g.*, Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig & Sendhil Mullainathan, *Human Decisions and Machine Predictions*, 133 Q.J. ECON. 237 (2018).

3. *See, e.g.*, State v. Loomis, 881 N.W.2d 749, 755 (Wis. 2016) (describing the use of COMPAS risk assessment by judge in determining sentence); Jennifer K. Elek, Roger K. Warren & Pamela M. Casey, *Using Risk and Needs Assessment Information at Sentencing*, NAT'L CTR. FOR ST. CTS. (2011), https://ncsc-search.squiz.cloud/s/redirect?collection=ncsc-meta&url=https%3A%2F%2Fwww.ncsc.org%2F__data%2Fassets%2Fpdf_file%2F0019%2F25174%2Frna-guide-final.pdf&auth=bIo81ujk6QRZWI0zqQO5bg&profile=_default&rank=1&query=using+risk+and+needs+assessment+at+sentencing (guiding judges and others involved in sentencing decisions on the use of risk assessment instruments).

4. *See, e.g.*, Kimberly Thomas & Paul Reingold, *From Grace to Grids: Rethinking Due Process Protections for Parole*, 107 J. CRIM. L. & CRIMINOLOGY 213 (2017).

5. *See, e.g.*, H. L. A. Hart, *Discretion*, 127 HARV. L. REV. 652, 662 (2013); Kent Greenawalt, *Discretion and Judicial Decision: The Elusive Quest for the Fetters That Bind Judges*, 75 COLUM. L. REV. 359, 361 (1975).

within the criminal law field and press for needed reform. We refer to these individuals and institutions as "stakeholders" throughout—defined broadly to include legislators, oversight bodies, civilian-review boards, researchers, journalists, activists, those directly impacted by the decisions, and the general public.

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*.

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.

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.

The Recon Approach starts from a place of acknowledging that human decision-makers have value in our legal system which machine learning cannot replace.[6] 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.[7] 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.

The Recon Approach takes inspiration from others in the social sciences who analyzed patterns in legal decision-making that were then used by

6. *See infra* Section III.B.

7. *See, e.g.*, Jeffrey J. Rachlinski, Sheri Johnson, Andrew J. Wistrich & Chris Guthrie, *Does Unconscious Racial Bias Affect Trial Judges?*, 84 NOTRE DAME L. REV. 1195, 1197 (2009) (finding evidence of unconscious racial bias among trial judges).

stakeholders as a tool for change.[8] 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.[9] 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[10] and statutory reform.[11] 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.[12] 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.

This Article proceeds in six Parts. Part II fully describes the Recon Approach and provides an example of how it might be implemented in one particular legal context: parole-release decision-making in California. This example has been the focus of our early work to implement the Recon

8. *See, e.g.*, Andrew Gelman, Jeffrey Fagan & Alex Kiss, *An Analysis of the New York City Police Department's "Stop-and-Frisk" policy in the Context of Claims of Racial Bias*, 102 J. AM. STAT. ASS'N 813 (2007) (analyzing sample of records from police stops and finding police stopped Black and Latinx people at higher rate than white people); David Arnold, Will Dobbie & Crystal S. Yang, *Racial Bias in Bail Decisions*, 133 Q.J. ECON. 1885 (2018) (analyzing court records and finding bail judges have bias against Black defendants).

9. *See, e.g.*, DAVID BALDUS, GEORGE WOODWORK & CHARLES PULASKI, EQUAL JUSTICE AND THE DEATH PENALTY: A LEGAL AND EMPIRICAL ANALYSIS 80–83 (1990).

10. *See* McCleskey v. Kemp, 481 U.S. 279, 287 (1987) (recognizing that Baldus study showed racial disparity in imposition of death penalty and holding that the evidence did not establish a violation of the Eighth or Fourteenth Amendments); John H. Blume & Sheri Lynn Johnson, *Unholy Parallels Between Mccleskey v. Kemp and Plessy v. Ferguson: Why Mccleskey (Still) Matters*, 10 OHIO ST. J. CRIM. L. 37, 56 (2012) (describing decades of impact litigation that built on the Baldus study and *McCleskey*).

11. *See, e.g.*, Robert P. Mosteller, *Responding to Mccleskey and Batson: The North Carolina Racial Justice Act Confronts Racial Peremptory Challenges in Death Cases*, 10 OHIO ST. J. CRIM. L. 103, 104 (2012) (describing enactment of North Carolina Racial Justice Act as response to *McCleskey* and study of death penalty decisions in North Carolina); Alex Lesman, *State Responses to the Specter of Racial Discrimination in Capital Proceedings: The Kentucky Racial Justice Act and the New Jersey Supreme Court's Proportionality Review Project*, 13 J.L. & POL'Y 359, 376 (2005) (same, for Kentucky).

12. *See, e.g.*, David Baldus, *When Symbols Clash: Reflections on the Future of the Comparative Proportionality Review of Death Sentences*, 26 SETON HALL L. REV. 1582, 1586 (1996) (describing cases applying various methods of comparative proportionality review).

Approach. Part II also explains how the Recon Approach can extend to sentencing and a variety of other contexts in which an adjudicator presides over a hearing and makes a discretionary decision.

Part III contrasts the Recon Approach with the Predictive Approach. Our objective is not to replace the Predictive Approach or deter its progress but rather to point the way to 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.

Part IV 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. Part V explains why development of NLP technology is integral to the long-term success of the Recon Approach. Parts VI and VII, respectively, discuss the technological challenges and the political challenges which need to be overcome in order to successfully execute the Recon Approach.

II. PILOTING THE RECON APPROACH IN THE CONTEXT OF PAROLE DECISIONS

To demonstrate more detail about the Recon Approach and its toolkit, this Part provides an example of early work to apply it in the context of parole-release decision-making in California. This Part also provides background about parole-release decisions and prior research in the area, and then describes development of a Recon Toolkit for this context. This example illustrates how the Recon Approach can provide guidance in many other contexts, provided they meet certain criteria and that both reconnaissance and reconsideration are critical for fulfilling its purpose.

A. BACKGROUND ON PAROLE HEARINGS AND PRIOR RESEARCH

Each year, the California Board of Parole Hearings (the Board) holds approximately 6,000 parole hearings for people in California prisons.[13] The purpose of the hearing is for the Board to decide whether a given individual who has served enough time to be eligible for release on parole (hereinafter

13. *See* CALIFORNIA BOARD OF PAROLE HEARINGS, CY 2019 SUITABILITY RESULTS, https://www.cdcr.ca.gov/bph/2019/10/24/cy-2019-suitability-results/ (last visited Apr. 28, 2021). In 2019, California scheduled 6,061 parole hearings that resulted in 1,184 grants of parole.

"parole candidate") is suitable for release.[14] 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."[15]

Parole hearings are generally overseen by one commissioner of the Board and a deputy who assists the commissioner.[16] 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.[17] 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.[18] The Board has broad discretion to decide whether a candidate is suitable for release and must produce publicly available transcripts from each hearing.[19]

The decision made at the hearing is subject to review by the Board's internal administrative review unit as well as California's Governor.[20] 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.[21] If a parole candidate is found unsuitable for parole, the opportunities to reconsider the decision are very limited. A

14. 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" 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 in note 20, a person may be found suitable for parole at the hearing but nevertheless not be granted release if the decision is later reversed.

15. *See* CAL. PENAL CODE § 3041(a)(2) (West 2018); *In re* Lawrence, 190 P.3d 535, 560 (Cal. 2008).

16. *See* California Board of Parole Hearings, Parole Consideration Transcripts (2007–2018) (35,105 transcripts on file with authors).

17. *See id.*; *see also* Kristen Bell, *A Stone of Hope: Legal and Empirical Analysis of California Juvenile Lifer Parole Decisions*, 54 HARV. C.R.-C.L. L. REV. 455, 472–73 (2019). 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. *Id.*

18. *Id.* 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. CAL. PENAL CODE § 3041.5 (West 2016).

19. CAL. PENAL CODE § 3042 (West 2017); *In re* Bode, 88 Cal. Rptr. 2d 536, 539 (Cal. Ct. App. 1999).

20. *See* CAL. PENAL CODE § 3041(b)(2) (West 2018) (authorizing the Board to review and reverse decisions); CAL. CONST. art. V, § 8 (authorizing the Governor to reverse decisions in murder cases, and to recommend that the Board change its decisions in non-murder cases).

21. *See* Interview with staff members who assist Gavin Newson in review of parole decisions, in Sacramento, Ca. (May 13, 2019).

parole candidate can request review by the Board's administrative review unit[22] as well as judicial review, but there is no right for appointed counsel to do so.[23] 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.[24] In practice, almost all candidates who are denied parole will remain incarcerated for years until the next opportunity for a parole hearing arises.[25] The wait can last from three years up to fifteen years long.[26]

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.[27] 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.[28] 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.[29] 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."[30] 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."[31] Consistency requires treating fittingly similiar cases alike, but what makes one parole candidate relevantly like (or unlike) another?

---

22. *See* CAL. 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).

23. *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.").

24. *See In re* Shaputis II, 265 P.3d 253, 267–68 (Cal. 2011).

25. *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)).

26. *See* CAL. PENAL CODE § 3041.5 (West 2016).

27. *See* Bell, *supra* note 17, at 480.

28. *See* California Board of Parole Hearings, Parole Consideration Transcripts (2002–2019) (35,105 transcripts on file with authors).

29. *See* CAL. CODE REGS. tit. 15, § 2402 (2001).

30. CAL. CODE REGS. tit. 15, § 2402(c)(1)(D) (2001).

31. CAL. CODE REGS. tit. 15, § 2402(d)(9) (2001).

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.[32] The sample size limits investigation to a small set of variables, ranging from fourteen to twenty-one variables.[33] 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.[34] In the meantime, changes in legislation and administrative regulations make the studies less directly applicable to current decision-making.[35]

B. PILOTING THE RECON APPROACH

In a pilot of the Recon Approach, we have begun creating a Recon Toolkit that includes tools designed primarily for reconnaissance and reconsideration of parole decisions. Through public records act requests and a lawsuit, we have acquired 35,105 parole hearing transcripts from 2007–2019 as well as other

32. *See* Bell, *supra* note 17, at 459 (studying sample of 426 parole transcripts in California); Beth Caldwell, *Creating Meaningful Opportunities for Release: Graham, Miller, and California's Youth Offender Parole Hearings*, 40 N.Y.U. REV. L. & SOC. CHANGE 245, 268 (2016) (studying sample of 107 parole transcripts in California); David R. Friedman & Jackie M. Robinson, *Rebutting the Presumption: An Empirical Analysis of Parole Deferrals Under Marsy's Law*, 66 STAN. L. REV. 173, 190 (2014) (studying sample of 103 parole transcripts in California); Kathryne M. Young, Debbie A. Mukamal & Thomas Favre-Bulle, *Predicting Parole Grants: An Analysis of Suitability Hearing for California's Lifer Inmates*, 28 FED. SENT'G REP. 268, 271 (2016) (studying sample of 754 parole transcripts in California). There are approximately 6,000 parole hearings held annually. *See* CALIFORNIA BOARD OF PAROLE HEARINGS, *supra* note 13.

33. *See* Bell, *supra* note 17, at 499 (considering sixteen variables in regression analysis on parole hearing decisions); Caldwell, *supra* note 32 at 275 (considering fourteen variables considered in regression analysis); Friedman & Robinson, *supra* note 32 at 195 (considering sixteen variables considered in regression analysis); Young et al., *supra* note 32, at 273 (considering twenty-one variables in regression analysis on parole hearing decisions).

34. *See* Bell, *supra* note 17, at 460 (being published five years after hearings began); Caldwell, *supra* note 32, at 245 (being published two years after hearings occurred); Friedman & Robinson, *supra* note 32, 189 (being published three years after hearings occurred); Young et al., *supra* note 32, at 271 (being published about six years after hearings occurred).

35. During the time when analysis was ongoing for the studies authored by Friedman and Robinson and Young, Mukamal, and Favre-Bulle, 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 and Caldwell, 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 PASSED REGULATORY PACKAGES, https://www.cdcr.ca.gov/bph/statutes/reg-revisions/ (last visited Apr. 28, 2021).

data that is not stated in the transcripts.[36] This other data includes the race/ethnicity of the parole candidate and whether the candidate was represented by a state-appointed attorney.[37]

The first step in reconnaissance is developing an information-extraction tool that uses Natural Language Processing (NLP) to review tens of thousands of transcripts. The tool will be trained to automatically extract information to answer about fifty questions, such as, "Which rehabilitation programs did the parole candidate participate in?" and, "If the candidate was written up for violating disciplinary rules in prison, what was the date of the last write-up?"

Next, another tool will be constructed to show what factors influence parole-suitability decisions and the relative influence of those factors. The tool will be based on information extracted from transcripts as well as other information not contained in transcripts, such as the parole candidate's race and whether the parole candidate's attorney was privately retained.[38] The model design will be user-friendly for stakeholders and adaptable over time. Stakeholders will be able to query the data for factors of their interest in response to the changing social and legislative landscape. For example, a stakeholder could run a query investigating how Black parole candidates fare relative to non-Black parole candidates when factors like the underlying crime, time-served, age, education-level, and history of prison misconduct are held constant. Figure 1 provides a snapshot of a preliminary reconnaissance visualization and data inspection tool that was built using data extracted from a sample of parole hearing transcripts.[39]

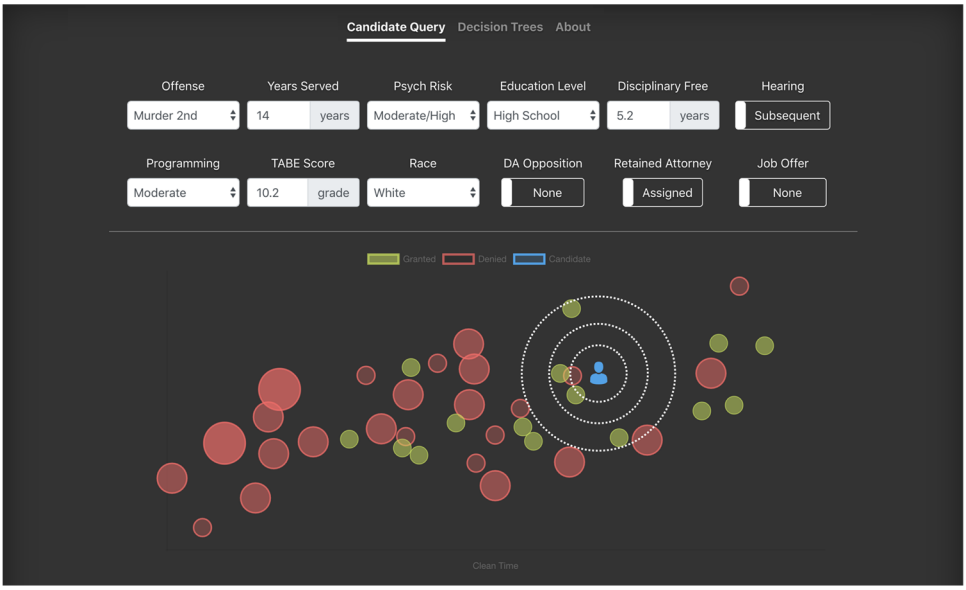
36. *See* Verified Petition for Writ of Mandate Ordering Compliance with the California Public Records Act, Voss v. California Dep't Corr. Rehab., No. CPF-20-517117 (Cal. Super. Ct. June 12, 2020).

37. Data is on file with authors.

38. The tool will also take into account the extraction noise in its modeling, similar to the way a social scientist would take into account the inter-rater reliability of her annotators when designing a model.

39. The data used to build Figure 1 was manually extracted from 426 transcripts and other information from youth offender parole hearings in California in 2014–2015. The same dataset was analyzed in the study conducted by Bell, *supra* note 17. In future development of reconnaissance tools, data will be extracted using NLP tools from 35,105 transcripts on file with authors.

Figure 1: Reconnaissance Tool Using Nearest Neighbors



The tool in Figure 1 shows how an imaginary candidate compares to actual cases that are relatively similar. To see this information, a 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 imaginary candidate is "plotted" as an individual icon amid circles that represent actual cases. Lighter green circles illustrate cases where parole was granted, and darker red circles illustrate cases where parole was denied. The size of the darker 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 individual icon are most similar to the imaginary candidate. Dotted rings around the individual icon show which circles would be considered "nearest neighbors" with more restrictive definitions of "near"—in essence, only looking at very similar cases.

In addition to reconnaissance tools, we are developing tools for the reconsideration of individual cases. Our goal is to create tools that identify cases that appear anomalous relative to general patterns in decision-making and flag those cases for reconsideration. As a hypothetical example, assume that in 90% of cases where a candidate has served over 25 years, has completed over 15 rehabilitation programs, and has no disciplinary write-ups in the last 5

years, the decision-makers find the candidate suitable for parole. Having identified the pattern, the reconsideration tool can identify the 10% of cases that are anomalous in the sense of having this same combination of factors, but nevertheless resulting in a denial of parole. The tool can flag these cases as anomalies warranting a second look. The technology needed to create such a reconsideration tool is not yet fully developed and some of the technological challenges are discussed in Part VI.

Cases flagged for reconsideration could receive a second look from various bodies such as the Board's administrative review unit, the Governor's review unit, or even appellate attorneys seeking to challenge denials. The reconsideration tool itself is agnostic with respect to who does the review of anomalous cases; that is, the tool itself does not designate who is best positioned to conduct the second look. The tool aims simply to provide an opportunity for a second look to happen when limited resources would otherwise prevent that from happening. After the second look occurs, the tool would be designed to receive feedback about which of the cases it flagged were actually reversed. The tool could then use this feedback to flag future cases that have similar features.

C. 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.)[40] 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 engrained, 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.[41] 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.

---

40. *See, e.g.*, Joss Greene & Isaac Dalke, *"You're Still an Angry Man": Parole Boards and Logics of Criminalized Masculinity*, THEORETICAL CRIMINOLOGY 1, 3–4 (2020) (discussing complexity and mixed results of quantitative analysis of race and parole decisions); Mindy S. Bradley & Rodney L. Engen, *Leaving prison: A Multilevel Investigation of Racial, Ethnic, and Gender Disproportionality in Correctional Release*, 62 CRIME & DELINQ. 2 (2016) (finding racial disparity in time-served prior to parole-release); Beth M. Huebner & Timothy S. Bynum, *The Role of Race and Ethnicity in Parole Decisions*, 46 CRIMINOLOGY 907, 925–26 (2008) (same); Bell, *supra* note 17, at 499 (finding Black candidates more likely to be denied at California youth offender parole hearings); Young et al., *supra* note 32, at 272 (not finding that race has statistically significant impact on California parole decisions); Stéphane Mechoulan & Nicolas Sahuguet, *Assessing Racial Disparities in Parole Release*, 44 J.L. STUD. 39 (2015) (not finding that race has statistically significant impact on parole decisions using national sample).

41. Many statistical tools from the Fairness in Machine Learning literature, such as calibration, propensity score weighting, or predicting on subgroups, could be used to develop an anomaly-detection tool that helps improve consistency as well as reduction in racial inequity.

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.[42] 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.[43]

The goal of the Recon Approach is not to identify the causal root of problematic patterns or assign blame. 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.

D. THE SCOPE OF 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,[44] 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.[45]

---

42. *See* Rachlinski et al., *supra* note 7, at 1197 (finding evidence of unconscious racial bias among trial judges).

43. *See* Bell, *supra* note 17, at 500.

44. *See infra* Section III.B

45. If a variable that is hypothesized to be relevant to the decision is not included, the resulting analysis will be vulnerable to omitted variable bias. *See generally* Hal J. Singer & Kevin W. Caves, *Applied Econometrics: When Can an Omitted Variable Invalidate A Regression?*, 17 ANTITRUST SOURCE 53 (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[46] 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.[47] 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.

III. DISTINGUISHING THE RECON APPROACH FROM THE PREDICTIVE APPROACH

Many technologists who are developing machine learning tools for use in criminal law use the Predictive Approach.[48] The Predictive Approach, broadly construed, aims to develop machine learning tools to predict a specified future outcome. This Part contrasts the Recon and Predictive Approaches by first summarizing the uses and critiques of the Predictive Approach before explaining the distinct potential of the Recon Approach.

46. *See, e.g.*, Roland G. Fryer J., *An Empirical Analysis of Racial Differences in Police Use of Force*, 27 J. POL. ECON. 1210 (2019).

47. *See* David Freeman Engstrom & Daniel E. Ho, *Algorithmic Accountability in the Administrative State*, 37 YALE J. REG. 800, 800–01, 809–15 (2020).

48. *See generally* Emily Berman, *A Government of Laws and Not of Machines*, 98 B.U. L. REV. 1277, 1280 (2018) (providing that predictive analytics are a primary focus of efforts to harness machine learning in criminal law).

A. THE PREDICTIVE APPROACH

Efforts to harness machine learning for use in criminal law have focused on making predictions about future outcomes in two areas: predictive policing tools and risk assessment instruments.[49] Predictive policing tools purport to identify individuals who are more likely to commit crime or geographic areas where crime is more likely to occur.[50] 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."[51] Approximately seventy percent of police agencies in the United States plan to deploy or increase use of predictive policing technology in the next two to five years.[52]

Actuarial risk assessment tools purport to estimate the degree of risk that a given individual poses for future violent behavior. The 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.[53] 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.[54] The basic approach began with statistical models in the 1920s,[55] 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.[56] Currently, criminal law practitioners across the United States use over sixty different risk assessment instruments across various

49. *See, e.g.*, Elizabeth E. Joh, *Feeding the Machine: Policing, Crime Data, & Algorithms*, 26 WM. & MARY BILL RTS. J. 287, 290 (2017) (describing use of predictive technology by police departments).

50. *See* Lindsey Barrett, *Reasonably Suspicious Algorithms: Predictive Policing at the United States Border*, 41 N.Y.U. REV. L. & SOC. CHANGE 327, 335 (2017).

51. *See* Joh, *supra* note 49, at 290–91, 298 n.73 (2017).

52. *See* William S. Isaac, *Hope, Hype, and Fear: The Promise and Potential Pitfalls of Artificial Intelligence in Criminal Justice*, 15 OHIO ST. J. CRIM. L. 543, 546 (2018).

53. *See* Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 811 (2014).

54. *See id.* at 813.

55. *See* Ariela Gross, *History, Race, and Prediction: Comments on Harcourt's Against Prediction*, 33 L. & SOC. INQUIRY 235, 236 (2008) (citing Clark Tidbitts, *Success or Failure on Parole Can Be Predicted*, 22 J. CRIM. L. & CRIMINOLOGY 11 (1931)).

56. *See* Sarah L. Desmarais & Samantha A. Zottola, *Violence Risk Assessment: Current Status and Contemporary Issues*, 103 MARQ. L. REV. 793, 813 (2020); *see generally* Shara Tonn, *Can AI help judges make the bail system fairer and safer?*, STAN. MAG. (Mar. 19, 2019), https://engineering.stanford.edu/magazine/article/can-ai-help-judges-make-bail-system-fairer-and-safer.

adjudicatory contexts.[57] 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.[58] In addition, parole board members rely on risk assessment scores in deciding whether to grant people release from prison.[59]

Critics of the Predictive Approach have argued that predictive policing tools and risk assessment instruments are not as accurate as they claim to be,[60] perpetuate racial bias,[61] and lack adequate transparency.[62] Proponents of the Predictive Approach continue working to address these criticisms.[63] Proponents also argue that human decision-makers fare no better than

57. Anna Maria, Barry-Jester, Ben Casselman & Dana Goldstein, *The New Science of Sentencing*, MARSHALL PROJECT (Aug. 4, 2015), https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing#.0olyDmAax.

58. *See, e.g.*, State v. Loomis, 371 Wis.2d 235, 243 (2016) (describing the use of risk assessments in sentencing); Megan Stevenson, *Assessing Risk Assessment in Action*, 103 MINN. L. REV. 303, 320 (2018) (describing the use of pretrial risk assessment at pre-trial detention decisions).

59. *See* Ebony L. Ruhland, Edward E. Rhine, Jason P. Robey & Kelly Lyn Mitchell, *The Continuing Leverage of Releasing Authorities: Findings from a National Survey*, 23–24, https://robinainstitute.umn.edu/publications/continuing-leverage-releasing-authorities-findings-national-survey.

60. *See, e.g.*, Michael Tonry, *Predictions of Dangerousness in Sentencing: Deja Vu All Over Again*, 48 CRIME & JUST. 439, 450 (2019) (describing meta-analyses which "conclude that positive predictions of future violence are too inaccurate to be used in sentencing"); Julia Dressel & Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, 4 SCI. ADVANCES 1, 3 (2018), https://advances.sciencemag.org/content/4/1/eaao5580/tab-pdf (showing that a widely used risk assessment tool is no more accurate at predicting than people with little or no criminal justice expertise).

61. *See, e.g.*, Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218 (2019) (summarizing arguments that algorithms in criminal justice perpetuate racial bias due to bias in input data and algorithmic methodology and arguing that the nature of prediction itself perpetuates bias).

62. *See, e.g.*, Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343 (2018) (explaining that many risk assessment instruments are deemed proprietary information and that the for-profit companies which develop them generally do not disclose the underlying datasets or the algorithms they use); Katherine J. Strandburg, *Rulemaking and Inscrutable Automated Decision Tools*, 119 COLUM. L. REV. 1851, 1862 (2019) (providing that many risk assessment instruments are built as opaque boxes in the sense that the patterns the instruments find in data are not explainable even to those who initially developed the software).

63. *See, e.g.*, Richard Berk, *Accuracy and Fairness for Juvenile Justice Risk Assessments*, 16 J. EMPIRICAL LEGAL STUD. 175, 184 (2019) (summarizing technical proposals to remedy bias in risk assessment algorithms); Hannah Bloch-Wehba, *Access to Algorithms*, 88 FORDHAM L. REV. 1265, 1272–73 (2020) (explaining how public records law can be used to access data about risk assessment instruments); Ashley Deeks, *The Judicial Demand for Explainable Artificial Intelligence* 119 COLUM. L. REV. 1829, 1833–34 (2019) (describing the need for developing explainable AI and progress toward that goal).

algorithms with respect to accuracy, bias, or transparency.[64] 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.

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.[65]

B. THE DISTINCT POTENTIAL OF THE RECON APPROACH

In presenting the distinct potential of the Recon Approach, it is helpful to draw upon the distinction between equitable justice and codified justice.[66] 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. Equitable justice requires discretionary moral judgment, which facilitates a case-by-case approach. Decisions are deemed fair insofar as they are justified on what are taken to be morally legitimate reasons.[67] Codified justice, on the other hand, refers to standardized application of specifiable rules.[68] The aim is to make the outcome of a decision determinable solely on the basis of rote application of a rule, thus pushing out discretionary judgment entirely.

Both types of justice have value in a legal system. Codified justice tends to diminish the vices of discretion like arbitrariness and bias while increasing efficiency and consistency.[69] Equitable justice brings in the virtues of discretion, such as individualized attention to unique case factors and

64. *See* Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Cass R. Sunstein, *Discrimination in the Age of Algorithms* (Nat'l Bureau of Econ. Rsch., Working Paper No. 25548, 2019), http://www.nber.org/papers/w25548 [https://perma.cc/JU6H-HG3W].

65. As discussed above in note 8, social scientists over the past several decades have made strides in analyzing patterns in discretionary decision-making. To date, however, this type of research has yet to leverage artificially intelligent technologies on a substantial scale. The thrust of the Recon Approach is to spur on investment in such technologies.

66. *See* Richard M. Re & Alicia Solow-Niederman, *Developing Artificially Intelligent Justice*, 22 STAN. TECH. L. REV. 242, 252–55 (2019) (explaining distinction between equitable justice and codified justice and arguing that artificial intelligence will tend to promote codified justice at the expense of equitable justice).

67. *Id.* at 252–53.

68. *Id.* at 253–54.

69. *See id.* at 253.

explanations of the reasoning underlying each decision.[70] The Recon Approach is designed to protect the pursuit of equitable justice through the human exercise of discretion.

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).[71] The Predictive Approach aptly fits this established niche by working on cost-effective minimization of 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.[72] Paraphrasing Justice Marshall, the power to exercise discretion is also an invitation to discriminate.[73] 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

70. *See id.* at 254.

71. *See* David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U. CAL. DAVIS L. REV. 653, 674–75 (2017) (explaining that the first step in developing a machine learning algorithm is to define what is to be predicted and specify it as a measurable outcome variable).

72. *See* KENNETH DAVIS, DISCRETIONARY JUSTICE: A PRELIMINARY INQUIRY (1969); H. L. A. Hart, *Discretion*, 127 HARV. L. REV. 652, 662 (2013); Walton v. Arizona, 497 U.S. 639, 664–65 (1990) (Scalia, J., concurring in part).

73. Furman v. Georgia, 408 U.S. 238, 365 (1972) (Marshall, J., concurring) (" '[C]ommitting to the untrammeled discretion of the jury the power to pronounce life or death' . . . was an open invitation to discrimination." (quoting McGautha v. California, 402 U.S. 183, 207 (1971)).

that everyone gets the benefit of that same process, there is a formal sense in which people are receiving equal treatment.[74] 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.[75] 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."[76]

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 was 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.[77] Or it may be that the decision is unreasonable upon 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.

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.[78] 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.

---

74. *See* Paul Stancil, *Substantive Equality and Procedural Justice*, 102 IOWA L. REV. 1633, 1636 (2017).

75. *See, e.g.*, BALDUS, *supra* note 9 (writing that procedural protections reduced but did not eliminate racial disparity in imposition of the death penalty).

76. Daniel E. Ho, *Does Peer Review Work? An Experiment of Experimentalism*, 69 STAN. L. REV. 1, 81 (2017).

77. 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." Louis E. Goodman, *In Defense of Federal Judicial Sentencing*, 46 CALIF. L. REV. 497, 498 (1958).

78. *See, e.g.*, Aziz Z. Huq, *A Right to a Human Decision*, 106 VIRGINIA L. REV. 611, 653–80 (criticizing arguments that human discretionary judgment is morally necessary in law).

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,[79] legal scholars,[80] psychologists,[81] 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."[82]

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.[83] 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.[84] Human discretion, when functioning well, acts as a way to respect and balance these several (and sometimes competing) values to reach a reasonable judgment.[85] In contrast, insofar as reliance is placed exclusively on predictive technologies like risk assessment tools, only the value of predicting and preventing crime is

79. *See, e.g.*, Lockett v. Ohio, 438 U.S. 586, 606 (1978).

80. *See generally* Jerry L. Mashaw, *Administrative Due Process: The Quest for a Dignitary Theory*, 61 B.U. L. REV. 885 (1981) (arguing that respect for the value of dignity calls for a process that allows for people to be heard and meaningfully participate in decisions made about them).

81. *See generally* TOM R. TYLER, WHY PEOPLE OBEY THE LAW: PROCEDURAL JUSTICE, LEGITIMACY, AND COMPLIANCE (1990) (explaining that, when processes provide an opportunity to participate and be heard, people feel more respected in the process and afford greater legitimacy to those overseeing the process).

82. Cade Metz & Adam Satariano, *An Algorithm that Grants Freedom, or Takes It Away*, N.Y. TIMES, Feb. 6, 2020.

83. *See generally* Kristen Bell, *A Reparative Approach to Parole Release Decisions, in* RETHINKING PUNISHMENT IN AN ERA OF MASS INCARCERATION (Chris W. Surprenant ed., 2018) (describing multiplicity of values at stake in parole-release decisions).

84. *See* MODEL PENAL CODE § 1.02(2) (AM. LAW INST. 2019) (listing multiple purposes of sentencing including inter alia proportionality of punishment to the gravity of the offense, rehabilitation, deterrence, incapacitation, preservation of families, reintegration of offenders into the community, as well as eliminating inequities in sentencing across population groups, ensuring humane treatment, and increasing the transparency, accountability, and legitimacy of the sentencing system).

85. *See* H. L. A. Hart, *Discretion*, 127 HARV. L. REV. 652, 662–63 (2013) (defining discretion and explaining that its most apt use is in contexts where there is an indeterminacy of aim).

taken into account.[86] This value would be privileged not necessarily because it is any more important but because it is most easily quantifiable.[87] 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.

IV. DEFENSES AGAINST PERPETUATING EXISTING PROBLEMS WITH THE STATUS QUO

This Part 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.[88]

86. *See* BERNARD E. HARCOURT, AGAINST PREDICTION: PROFILING, POLICING, AND PUNISHING IN AN ACTUARIAL AGE 58 (2007).

87. *See id.* at 188.

88. *See* Engstrom & Ho, *supra* note 47; 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).

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 above,[89] 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

89. *See supra* Section II.C.

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.

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.[90]

V. THE IMPORTANCE OF DEVELOPING NATURAL LANGUAGE PROCESSING (NLP) TOOLS

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

90. *See* Engstrom & Ho, *supra* note 47, at 807 (calling this type of review "prospective benchmarking" and setting forth reasons why it would be valuable developing AI decision-making tools within administrative law).

NLP tools for the recon context.[91] Scholars have proposed this type of work-around as an alternative to NLP in other contexts.[92] 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.[93]

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.[94] The critical advantage of developing an NLP tool to

91. *See infra* Part VI.

92. *See* Engstrom & Ho, *supra* note 47, at 848 ("Agencies have deployed significant resources to use NLP techniques to convert unstructured text into structured data, but a first order solution—one that might in fact be cheaper in the long run—would be to standardize inputs.").

93. *See* CAL. PENAL CODE § 3041.5(4) (West 2016).

94. 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

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.[95] 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 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.[96] At a subsequent hearing, a different commissioner granted parole and stated that the candidate was not contesting an underlying aspect of the same offense.[97] 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

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.

95. *See* California Board of Parole Hearings, Parole Consideration Hearings 4, 36 (January 2015) (transcript on file with author); Graham Todd, Catalin Voss & Jenny Hong, *Unsupervised Anomaly Detection using Language Models*, Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science 66 (Nov. 20, 2020) (discussing how Recon Toolkit found an instance in which a parole board commissioner called the parole candidate a "smart ass").

96. *See* California Board of Parole Hearings, Parole Consideration Hearings 121 (February 2016) (transcript on file with author).

97. *See* California Board of Parole Hearings, Parole Consideration Hearings 215 (August 2017) (transcript on file with author).

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.[98] 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.[99] 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

98. *See* Rob Voigt, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky & Jennifer L. Eberhardt, *Racial disparities in police language*, 114 PROC. NAT'L ACAD. SCI. 6521 (2017).

99. *See, e.g.*, Ashish Vaswani et al., *Attention is all you need*, *in* 30 ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (2017) (introducing a neural network architecture, the Transformer, which improves on then-state-of-the-art Recurrent Neural Networks (RNNs) by providing a more effective memory of context and the ability to parallelize computation); Jacob Devlin, Ming-Wei Chang, Kenton Lee & Kristina Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (June 2019) (showing that BERT, an instantiation of the Transformer architecture, can be pre-trained on generic English-learning tasks and fine-tuned to specific tasks like translation, summarization, and generation); Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Darioa Amodel & Ilya Sutskever, *Language Models are Unsupervised Multitask Learners*, OpenAI Blog 1.8, 9 (2019), https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (reviewing BERT and other transformer models that are first pre-trained on generic English-learning and then fine-tuned to a specific task, and finding that the models perform well on each individual task); Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov & Quoc V. Le, *XLnet: Generalized Autoregressive Pretraining for Language Understanding*, *in* 32 ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 5754 (2019) (introducing XLNet which improves on BERT's performance on a range of NLP tasks); Tom B. Brown et al., *Language Models are Few-Shot Learners* (2020), https://arxiv.org/pdf/2005.14165v2 (showing that Transformer-based models can perform well when they are trained generally to understand English, with only a small fine-tuning operation at the end to learn to do any specific task).

feedback on whether the identified cases are indeed anomalies of potential interest or are simply red herrings.[100] 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 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.

VI. TECHNOLOGICAL CHALLENGES

This Part 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.

A. 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

100. *See* Todd et al., *supra* note 95.

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.[101] 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.[102] These techniques do not scale well to parole hearing transcripts which are approximately 10,000 words.[103]

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.

101. *See supra* note 99.

102. *See, e.g.*, *supra* note 99. Larger models like GPT-3 proposed by Brown et al., see *supra* note 99, can handle up to 2048 so-called "word-pieces" (also referred to as "tokens") which may cover up to 1,500 words of normal speech, but these models cannot yet be run by organizations with access to reasonable amounts of computing power. *See* RISHI BOMMASANI ET AL., STANFORD CTR. RSCH. ON FOUND. MODELS, ON THE OPPORTUNITIES AND RISKS OF FOUNDATION MODELS 11 (2021), https://crfm.stanford.edu/report.html.

103. *See generally* California Board of Parole Hearings, Parole Consideration Transcripts (2007–2018) (35,105 transcripts on file with authors). These transcripts produce on average 27,000 word pieces ("tokens") using the BERT encoding. *See* Devlin et al., *supra* note 99, at 4173.

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.[104]

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.

B. 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,[105] and other data that is not included in the text. Regression analysis is often used to perform this type of task.[106]

104. *See generally* Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov & Christopher D. Manning, *Hotpot QA: A Dataset for Diverse, Explainable Multi-hop Question Answering*, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing 2369 (2018), https://www.aclweb.org/anthology/D18-1259.pdf.

105. These statistics should include a measure of the reliability with which the NLP tool extracted the correct answers to its queries.

106. Regression analysis is a statistical technique used to understand the relationship between independent variables which are "thought to produce or be associated with changes

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.[107]

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.[108] 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 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

in [a] dependent variable." Daniel L. Rubinfeld, *Reference Guide on Multiple Regression*, *in* REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 303–57 (3d ed. 2011).

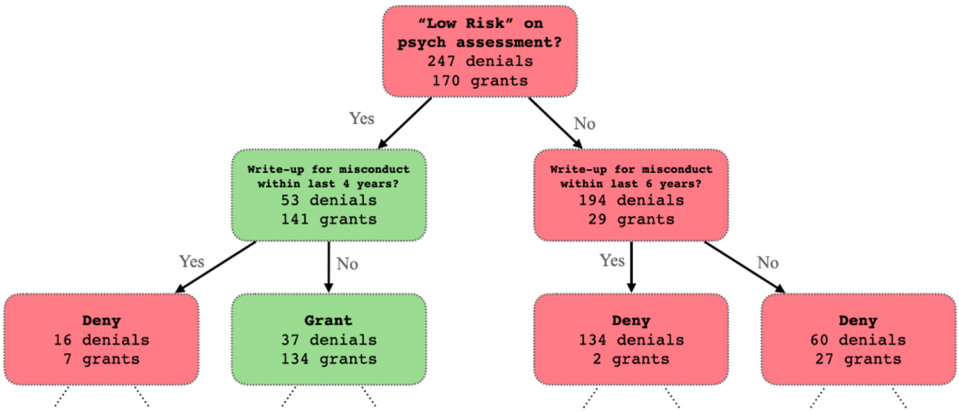
107. *See id.* at 320, 345 (explaining r-squared values as measure of fit and p-values as measure of statistical significance).

108. *See* TREVOR HASTIE, ROBERT TIBSHIRANI & JEROME FRIEDMAN, THE ELEMENTS OF STATISTICAL LEARNING: DATA MINING, INFERENCE, AND PREDICTION 10, 18 (2d ed. 2017).

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, a version of which is illustrated and described in Figure 1 above. 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 below in Figure 2.

Figure 2: Decision Tree Model of Parole Hearings



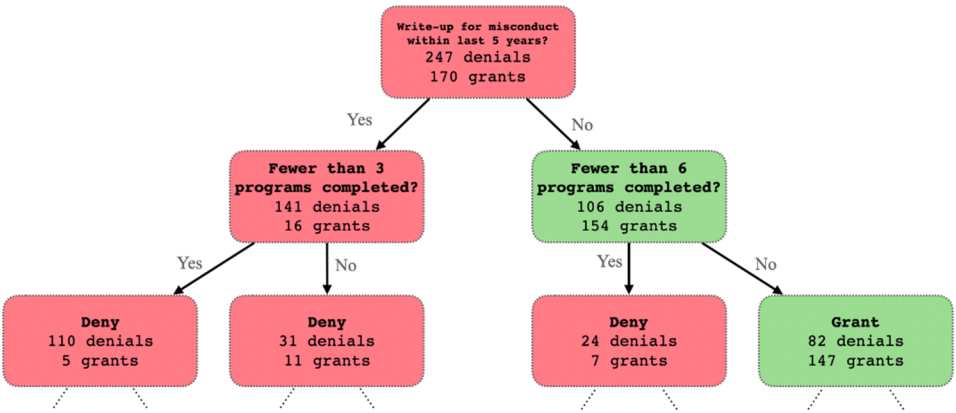
This figure illustrates an excerpt of a larger decision tree that was generated from a dataset extracted from a sample of parole transcripts in 2014–2015.[109] In this excerpt, only the top three levels of the tree are shown. The tree reads from the top down. At each step, the algorithm partitions the 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.

109. *See supra* note 39.

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.[110] To see why there is a need for a "fit" metric, consider Figure 3 which is built from the same sample of parole hearing decisions as Figure 2. It illustrates an alternative decision tree that was generated over the same set of transcripts as Figure 2. Again, as in Figure 2, this is an excerpt of a tree and bottom leaves are not shown.

Figure 3: Alternative Decision Tree Model of Parole Hearings



The primary criteria for sorting decisions in Figure 3 is whether or not a parole candidate received a disciplinary write-up within the last five years. In Figure 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

110. 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.

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?"[111] 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. Almost no metrics exist to help choose among multiple trees that predict equally well because, 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 more aptly describes a pattern of decisions than Figure 3 or some other tree that is generated randomly? Additional techniques are required to answer this question.[112] 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

111. Further, multiple trees are often combined to form powerful predictive algorithms, for example in Random Forest classifiers, dating back to the 1990s. *See* Tin Kam Ho, *Random decision forests*, Proceedings of Third International Conference on Document Analysis and Recognition (1995).

112. Naive permutation tests that are applicable to black box machine learning models more broadly can also be used to test the decision trees' robustness, but these lack well-defined null hypothesis and thus cannot be used for statistical significance testing.

been modifying regression models easily to handle such noise,[113] 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.

VII. POLITICAL CHALLENGES

This Part 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.

A. 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

113. *See* PAUL GUSTAFSON, MEASUREMENT ERROR AND MISCLASSIFICATION IN STATISTICS AND EPIDEMIOLOGY: IMPACTS AND BAYESIAN ADJUSTMENTS 12 (2003).

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.[114] We postponed our work for approximately nine months of negotiation which led to litigation about our right to access race data.[115] 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."[116]

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.[118] We also support development of independent commissions within state governments which are charged to collect and study criminal justice data; California has recently created such a commission.[119] Lastly, we encourage publication of the "non-

114. *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).

115. *See id.*

116. *See* Voss v. California Dep't of Corr. Rehab., No. CPF-20-517117 (Cal. Super. Jul. 16, 2020), https://www.eff.org/document/order-voss-v-cdcr; Brodheim v. California Dep't of Corr. Rehab., No. CPF-20-516978 (Cal. Super. Jul. 16, 2020), https://www.eff.org/document/order-brodheim-v-cdcr-voss-v-cdcr-companion-case.

117. *See* GA. CODE ANN. § 42-9-53 (West 2017).

118. *See* MEASURES FOR JUSTICE, https://measuresforjustice.org/ (last visited Apr. 28, 2021).

119. *See* CAL. GOV'T 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

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.

B. RESEARCHER-CAPTURE

The Recon Approach is potentially vulnerable to a phenomenon that administrative law scholars refer to as "regulatory capture" or "agency capture."[120] 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.[121]

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

information, and reasonable assistance in any matters of research requiring recourse to them, or to data within their knowledge or control").

120. *See, e.g.*, J. Jonas Anderson, *Court Capture*, 59 B.C. L. REV. 1543, 1555 (2018).

121. *See id.*

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 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.

VIII. 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!"[122] 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.

122. THOMAS MORE, UTOPIA 42 (Paul Turner ed., Penguin Books 2003) (1516).