Stanford University
Human-Centered
Artificial Intelligence

DECEMBER 2021

A New Direction for Machine Learning in Criminal Law

Kristen Bell, Jenny Hong,
Nick McKeown, Catalin Voss

THE AMERICAN CRIMINAL LEGAL SYSTEM is rife with—and perpetuates—inequality. These discrimination problems across racial, socioeconomic, and other lines are well-documented, but studying the problem is still a resource-intensive process. Combing through court transcript hearings, manually categorizing case files, and other investigative techniques require hours of time, extensive manual labor, and research dollars that many civic and other organizations simply lack. Technology may be able to relieve some of this burden.

Undoubtedly, many current applications of artificial intelligence (AI) technology have enabled, if not exacerbated, the discrimination problems in the criminal legal system. For instance, critics argue that because algorithms are trained on datasets reflecting centuries of racism (e.g., on arrest rates across racial groups), they tend to overestimate the risk of recidivism among defendants of color as compared to white defendants. There is justified concern about the use of technology in the criminal legal system and how it could focus on "technical fixes" for broad social and political problems in the criminal justice system and only make the problems worse.

KEY TAKEAWAYS

- Using machine learning to analyze decision-making in the criminal legal system could be a valuable way to identify discrimination and facilitate reconsideration of decisions where justice was inconsistently applied—but reconsideration is still a decision, and stakeholders in criminal law processes should consider whether and how machine learning should play a role in that decision.
- We propose a two-pronged "Recon Approach" to use machine learning in criminal law: *reconnaissance*, where machine learning identifies patterns in a human decision-making process, and *reconsideration*, where machine learning then focuses on individual decision cases.
- The Recon Approach is meant for scenarios where humans make discretionary decisions, there are records which describe decision factors, and those records are analyzable by a machine learning tool (e.g., a hearing transcript readable by a language-processing algorithm).
- Policymakers should consider implementing stronger, clearer public record laws to ensure researchers have access to the necessary data to conduct these reviews.

In our new paper, we propose using machine learning (ML) to analyze decision-making in the criminal legal system. Instead of using ML to assess those being put through the system, we argue for using ML to analyze decisions that people in power have already made—thereby bringing increased transparency to those decisions, identifying patterns, and shining a light to help people see potential injustices. The aim is not to predict human behavior or replace human decision-making, but to better understand the factors that led to past decisions in the hopes of facilitating increased fairness and consistency in how criminal law is applied. We call it the "Recon Approach."

Introduction

Using ML to analyze decision-making in criminal law raises many ethical concerns. Extracting information from past decisions may be done in a way that produces data that is subsequently used to discriminate. The selection of which factors to weigh and ignore when a natural language processing (NLP) algorithm reads a legal document likewise risks skewing the analysis against already vulnerable populations. Constantly reassessing and addressing these issues is paramount to the Recon Approach. We believe if the Recon Approach is executed carefully, it can be used as a tool to help people identify and combat discrimination both at the systematic level and in individual cases.

We demonstrated the power of this approach by developing an early-stage "Recon Toolkit," which we applied to analyze 35,105 California parole hearing transcripts from 2007-2019 obtained via a public records request. The Recon Toolkit uses NLP to read all the transcripts and identify key factors raised in parole hearings (e.g., the offense, years served, psychological risk assessment score, year the last disciplinary infraction occurred, education level, whether or not the district attorney appeared at the hearing, etc.). The extraction of this data enables us to analyze the extent to which these factors explain parole decisions. We identify several mechanisms that introduce significant arbitrariness into the parole process. Factors outside of a candidate's control disproportionately explain parole decisions when controlling for relevant case factors. Non-white candidates are less likely to be represented by private attorneys, and non-white candidates also speak fewer words during the hearing.[1] Building on this analysis, software could be developed to visualize how factors impact parole decisions. Because our toolkit can identify relevant case factors for every transcript—not just a sample of transcripts like manual social science study could—it can also be used to comparatively investigate cases at an individual level. Researchers could build an imaginary parole candidate profile and compare that to relatively similar, real-world cases using a nearest-neighbor search. For instance, if two parole candidates were similar in terms of the underlying crime, time served, and history of prison conduct, but one was Black and the other was white, researchers equipped with the Recon Toolkit can investigate how the Black individual fares in that scenario.

Our paper discusses key areas where further research should be conducted in NLP, statistics, and ML ethics in order to move the Recon Approach forward. Researchers and technologists cannot explore these possibilities on their own; input from policymakers and the general public is critical.

1 We discuss these findings in detail in a forthcoming paper (in submission).

The Recon Approach

Reviewing decision-making within the criminal legal system for bias or other injustices is a complicated, intensive process. Depending on the scope of review, researchers might have to read thousands of hearing transcripts, each of which may be hundreds of pages long—requiring significant time and manual labor. Nevertheless, these analyses can lead to meaningful change in the justice system, as with major reviews of death penalty cases that identified systemic racism and led to crucial reforms.

Depending on the scope of review, researchers might have to read thousands of hearing transcripts, each of which may be hundreds of pages long—requiring significant time and manual labor.

Our paper proposes a solution to lower the reviews' cost: machine learning. Rather than make predictions about those going through the justice system, or try to make assessments helpful to judges or parole board members, we focused on analyzing justice system decisions, after the fact, for fairness and consistency problems. Those investigating criminal justice system decisions and pushing for reform, rather than decision-makers themselves, are the intended users of the Recon Approach.

The first step in the Recon Approach is "reconnaissance," which involves running NLP and other algorithms on transcripts of hearings in which a discretionary decision is made by an official. As the algorithms read over the transcripts—far more quickly than any human could—they can identify patterns in the text, such as which factors are raised in a hearing (e.g., age, prior offenses, extenuating circumstances). Researchers could also use algorithms to create "decision trees" based on this information to model the decision logic of a human decision-maker in the room (e.g., on a parole board).

The second step is "reconsideration," which takes the higher-level analysis from the reconnaissance portion and applies it to individual cases. Where reconnaissance focuses on identifying high-level trends, reconsideration focuses on identifying anomalous cases yielding decisions that differ from the overall patterns. For example, if reconnaissance shows that parole hearings with a district attorney present routinely leads to candidates with similar case factors being denied parole, that could be a reason to reconsider hearings with district attorneys in attendance. This could be done via techniques such as a "nearest neighbor" algorithm, where the reviewer defines the basis of the comparison. For instance, if two parole candidates are similar across underlying crime, time served, and history of prison misconduct, but one is Black and the other white, we could use the Recon Approach to investigate how the Black individual fares in that scenario.

We argue that reconsideration ought to go hand in hand with reconnaissance. Without reconnaissance, researchers would lack an analysis of broader trends across thousands of hearings; without reconsideration, researchers would have trouble narrowing in on specific incidents. Using reconsideration, without first applying reconnaissance, to analyze a single case means that decisions may seem consistent within a case but prove to be part of a consistently unfair trend among similar cases. Skipping the reconnaissance step risks perpetuating systemic unfairness in decision-making.

Policy Discussion

Our goal is not to swap out judges, juries, or parole boards for machine learning tools. Algorithms and other technologies cannot fully capture the deep, nuanced, social and emotional aspects of human-run legal hearings. But processing hundreds of thousands of pages of dialogue across decades, and then analyzing systemic trends, cannot be done by hand. We believe the Recon Approach could improve analysis of what human parole boards, judges, and others have already decided. The approach stands in contrast to other uses of ML that seek to influence decision-makers before they even reach a decision.

Our goal is not to swap out judges, juries, or parole boards for machine learning tools. Algorithms and other technologies cannot fully capture the deep, nuanced, social and emotional aspects of human-run legal hearings. But processing hundreds of thousands of pages of dialogue across decades, and then analyzing systemic trends, cannot be done by hand. We believe the Recon Approach could improve analysis of what human parole boards, judges, and others have already decided.

We propose three principles for ethically developing the Recon Toolkit—though subsequent work could identify more. First, a diverse group of stakeholders should select the "chosen factors" analyzed by an algorithm. This group should, at least, include the decision-makers themselves, those about whom decisions are made, those individuals' attorneys, legislators, researchers, and members of the general public. Second, any researcher implementing the Recon Toolkit must be transparent about and publish an updated list of all the factors they selected to use in the analysis process (e.g., race, time served, etc.), all those considered but not ultimately chosen, and background on the underlying relationships or potential relationships between factors. Third, if a reconsideration tool is deployed with the goal of

improving consistency in decision-making, researchers should periodically compare its outputs to those of a tool that randomly selects cases for review. If the randomly selected cases are overturned at the same rate or more often than those flagged by the researchers' algorithm, they should reassess their own tool.

Further, in order for a reconsideration tool to improve equity in decision-making, those working on the tool also need to consider their role in the specific decision-making context. For example, in the parole context, a reconsideration effort can only be used to increase parole grants but not to turn grants into denials. Currently, the state governor's office reviews all grants of parole and, upon their personal reconsideration, a significant portion of grants are changed to denials. Once a parole grant has been approved by the governor's office, the person is released from prison and there is no reconsideration. But there is no systematic review of the thousands of decisions to deny parole—a reconsideration tool for parole could help facilitate such a review. It is within this context that a reconsideration tool could effectively reduce mass incarceration.

To ensure researchers have ready access to the necessary data about decision-making in criminal law, policymakers might consider strengthening and further clarifying public record laws. The paper's initial findings drew on a dataset of over 35,000 California parole board hearing transcripts. Yet to gain access to additional and vital data, such as the race of a parole candidate, we had to file a lawsuit against the California Department of Corrections and Rehabilitation. We received help from the Electronic Frontier Foundation and were ultimately successful, with a court finding race data to be public record, and one with "weighty public interest in disclosure." The litigation, however, delayed the research by nearly two years. Many other research organizations

To ensure researchers have ready access to the necessary data about decision-making in criminal law, policymakers might consider strengthening and further clarifying public record laws.

will not have the time, money, or personnel to replicate this data acquisition process. Colleagues in other states may face even greater hurdles when acquiring their data: California is relatively unique in placing its parole hearing transcripts in the public record. Our experience points to a need to strengthen and clarify public record laws around the country to facilitate this type of research. We see reason for hope among nonprofit organizations like Measures for Justice that work to gather criminal justice data and make it available to the public.

Based on our findings, policymakers should also consider establishing independent commissions within states to collect and study data related to criminal law, and require agencies to publish "non-finding" notices whenever they deny researchers access to data. Coupled with bolstering public record laws, this would provide greater public transparency into agency data decisions. It would also force a reputational cost on any agency denying data to members of the research community.

Policymakers should consider if and how data automatically extracted from case records can assist in decision reviews. Simply tasking one individual or a single administrative unit with reviewing decisions may lead to imbalanced reviews. For example, in California, the governor reviews parole board decisions but has limited resources to do so. In practice, the governor reviews all parole grants but only a tiny fraction of the thousands of decisions denying parole. In conjunction with legislative changes to a decision-making process, policymakers may want to explore instituting reviews of historical case denials through specified criteria.

We propose a novel ML toolkit for systematically analyzing data, at scale, about how state officials make decisions that alter the lives of people going through the criminal legal system. The Recon Toolkit could reduce the time, money, and resources needed to process decision transcripts by hand, and thus greatly reduce the barrier to researchers and other individuals looking to analyze the system through a lens of fairness and consistency. It is through technology and policy interventions acting in concert that ML can hopefully advance a more transparent understanding of how power is exercised in criminal law.

The original article, **"The Recon Approach: A New Direction for Machine Learning in Criminal Law,"** can be accessed at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3834710.

Stanford University's Institute on Human-Centered Artificial Intelligence (HAI), applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu.**

Kristen Bell is an assistant professor at University of Oregon School of Law.

Jenny Hong is a Ph.D. student in the department of management science & engineering at Stanford University.

Nick McKeown is a professor of computer science and electrical engineering at Stanford University.

Catalin Voss is a Ph.D. student in computer science at Stanford University.

Stanford University
Human-Centered
Artificial Intelligence

Stanford HAI: Cordura Hall, 210 Panama Street, Stanford, CA 94305-4101

T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**