

Do Judges Behave Like Algorithms?

Riya Manchanda^{*1}, Eric Chen^{*1}, Chloe Zhu¹, Cynthia Rudin¹, Brandon Garrett¹, Songman Kang²

¹Duke University

²Sungkyunkwan University

riya.manchanda@duke.edu, eric.y.chen@duke.edu, qinyu.zhu@duke.edu, cynthia@cs.duke.edu, bgarrett@law.duke.edu, songmankang@skku.edu

Abstract

What if judges already behave like algorithms? As artificial intelligence and algorithms are deployed in many settings, including the judicial system, many have debated whether judges should be allowed to rely on them. We suggest this debate about the role of algorithms in the justice system tracks a longstanding debate regarding the comparative benefits of judges relying on rules versus standards. Rules are clear, consistent, and predictable, but may not capture individual and equitable circumstances that more flexible standards rely on. We do not claim when or whether rules or standards are preferable. Instead, we ask whether judges follow predictable, algorithmic-like rules already. If judges already follow consistent, formula-like rules based on discrete and static factors such as criminal history, age, and charge type, then judicial behavior may be improved. However, if judges rely on individualized information that cannot be identified through court data, then standards-based decision-making may be more challenging to understand or improve. This work explores these questions by studying judicial decision-making in misdemeanor bail hearings in Harris County, Texas. Using available court data, we investigate whether magistrate judges follow what resembles an algorithm; whether they consider the same variables in their decision-making; and whether they are consistent with themselves and with each other. To do this, we train machine learning models for each judge, measure variable importance metrics to determine important variables for each judge’s decision-making, and analyze outcomes of similar cases for judges. Our results reveal that these judges generally behave algorithmically: their decisions can be captured by small, interpretable formulas. However, in some cases, judges differ substantially, leading to surprising inconsistency and unequal treatment across similar defendants. Identifying cases where algorithms do not explain judicial decision-making can improve the justice system by focusing attention on decisions where individualized standards, rather than rules, better explains outcomes.

Full Appendix and Public Code —

<https://github.com/eychen2/AreJudgesAlgorithms>

Introduction

Having an impartial judge decide one’s dispute is an essential element of due process, or fair trial rights, around the

world (Garrett 2025; Deeks 2019; Crootof et al. 2023). An important dimension of due process is having a judicial officer who makes these decisions in a reasoned and consistent manner. Yet, there has been a longstanding tension between making decisions based on *rules*, which can lead to consistent and predictable outcomes, and the use of more discretionary *standards*, which can provide more equitable outcomes based on individual, case-specific considerations (Schlag 1985). There are important questions concerning when and whether rules versus standards are desirable, and there are tradeoffs concerning the choice of which to employ in legal decision-making (Kaplow 1992). For more information about rules and standards, refer to Appendix A. In recent years, the debate about the use of algorithms and AI in the justice system has tracked many of the same concerns expressed in the longstanding debate between the comparative desirability of rules versus standards. One side argues that many algorithms are reductive and fail to account for the case-specific factors that a human judge must consider (Harcourt 2015). This concern is particularly pressing in areas where equity and a judge’s discretionary moral judgment are involved (Re and Solow-Niederman 2019). Others note that human judges are also prone to certain types of inconsistencies and biases in their decision-making (Rachlinski and Wistrich 2017; Eaglin 2023). The type of data and the type of judicial task may also affect the choice of whether algorithms or discretionary judging are more desirable. Using AI or algorithms is more advantageous if prediction is straightforward and rule-based, while an element of human judgment will be more advantageous if omitted variables make accurate prediction more difficult (Fagan and Levmore 2019). Highly relevant to this debate is whether judges *already* behave like algorithms and follow rules, or rather, whether they tend to decide cases based on more case-specific standards. Let us define this question more carefully. By an “algorithm,” we refer to a simple formula, perhaps consisting of a few understandable rules. This simple formula involves only a few additions, subtractions, and multiplications. We call such formulas “interpretable” or “understandable,” because the actual factors and weights they rely upon are provided (Garrett and Rudin 2024a). We are not referring to “black boxes,” i.e., complicated or proprietary mathematical formulas, since we believe those should be avoided in criminal justice decision-making, as they raise

^{*}These authors contributed equally.

serious constitutional concerns and have not been shown to be better than interpretable models (Garrett and Rudin 2024a,b). The question we ask here is when and whether judges act like simple formulas. Are judges consistent in their judgments the same way a simple formula (or rule) would be? If so, can we identify these formulas or rules? If instead they employ standards, can we identify them in order to shed light on how they make their decisions? These questions are central to debates on whether algorithms should be used in the justice system. If judges already follow consistent, formula-like rules based on quantifiable cues such as prior offenses, age, and charge type, then judicial behavior could be understood and even improved through interpretable models rather than being replaced by them. Conversely, if their decisions deviate widely from any learnable rule, then the challenge is to identify these more flexible standards, assess the consistency of decision-making, and, in turn, ensure that the use of those standards can be understood and even improved.

We examine this question by focusing on data from the third-largest urban county in the United States, Harris County, Texas, which includes the city of Houston. Our data is from the period after the 2019 entry of the O'Donnell Consent Decree, a landmark reform in Harris County designed to eliminate wealth-based disparities in bail decisions, and gave magistrates greater discretion in their cases, particularly regarding a smaller number of more serious misdemeanor cases.

In this work, we examine the decision-making by magistrates who are hired by the County to report to the elected judges and to make initial decisions regarding the comparatively more serious misdemeanor cases. Specifically, we focused on the cases in which the magistrates decide, after a bail hearing, during which they hear from the district attorney and public defender, whether to recommend that a person be given a personal bond and released, perhaps with supervision conditions, or whether a person should be given a secured bond with a cash bail amount that they would have to pay to be released, possibly also with supervision conditions if they are able to pay or obtain a bond for that amount. Given the size of the jurisdiction, there are tens of thousands of these bail hearings and magistrate decisions each year. Our unique dataset contains over 22,000 unique cases decided by 21 magistrates.

To clarify, this work does not advocate replacing judges with algorithms. An essential concern is *recourse*: humans are always needed, including when an algorithm is wrong, when cases raise individual or extenuating circumstances, or in cases where a rule-based decision would be appropriate, but the data about the individual is incorrect and the algorithm's input needs to be adjusted (Garrett and Rudin 2024a). Regarding the latter situation, while cases with data errors should be rare, they are not in practice, partly due to pervasive errors in administrative criminal justice data (Pepper, Petrie, and Sullivan 2010). The first type of case – where equitable discretion or standards should be used – raises a normative question about the desirability of using equitable, standards-based approaches. Without addressing that normative question, we instead ask whether judges act

like algorithms generally, including whether they are consistent with each other; whether they are needed mainly to intercept unusual cases; and whether there is some category of individualized decision-making that they already engage in. Examining the potentially algorithmic nature of judicial reasoning – and the types of cases that judges do not view as suitable for rule-based approaches – opens a path toward improved decision making and greater accountability.

Summary of experiments and results: To investigate our first question – “**When do judges not follow any reasonable algorithm, and are there circumstances leading to these inconsistencies?**” – we generated the “Rashomon set” of every possible good model of a flexible, nonparametric function class, namely, sparse decision trees. Each of these models represents a possible simple decision function that matches the judge's decisions as accurately as possible. For each magistrate, we identified cases that were misclassified for every function in the Rashomon set; these “unexplainable cases” are judicial decisions where no good model can predict the judge's ruling correctly.

Unexplainable cases amount to 2%-33% of each judge's cases. Manual analysis of a sample of hearing records and probable cause forms revealed that 100% of the sample of these unexplainable cases arose from either: (1) data-entry or form errors, (2) information about charges outside Harris County that were not represented in the dataset, or (3) case-specific qualitative details that judges referenced but were absent from structured data. These findings suggest that even if much of judicial reasoning were algorithmic, real-world inconsistencies often stem from gaps in information and record-keeping rather than deliberate arbitrariness. These results suggest that there are systematic gaps in record-keeping that could be mitigated.

Our second question is “**Can each judge be described by a simple algorithm?**” We attempted to model each judge with a simple decision tree and reported the degree to which each magistrate's bail decisions can be represented by such rules (disregarding the unexplainable cases, which arise from a noisy generative process rather than a deliberate one). We found that across thousands of cases, most judges could be approximated by small decision trees (depth ≤ 5) using features such as prior revocations, pending cases, and defendant age, achieving accuracies between 75% and 100%, where at least 85% of the decisions for each judge can be algorithmically determined for 17 of the 21 judges. This indicates that a substantial portion of judicial behavior reasonably follows systematic patterns that can be summarized in a few simple conditions, although some judges behaved more algorithmically than others.

Digging further, we performed a variable importance analysis to answer: “**Do judges consider the same variables in decision making?**” Most variable importance analyses are flawed since they only consider variable importance for one model, despite the existence of the Rashomon set – the set of good models for the data – which often contains many equally accurate models that use completely different variables. Our analysis, instead, averages variable importance over good models to identify variables that are generally important. We find that there are three generally im-

portant variables across judges – age, warrantct (number of previous warrants), and pre-defined flags of specific types of crime – but the degree to which variables are important and the choice of additional important variables differ substantially between judges. Thus, even if each judge behaves algorithmically, they act like totally different algorithms.

Finally, we delve deeper into the differences between judges, asking: **“Are judges consistent with each other? When do they disagree?”** To do so, we test whether a model trained to predict one judge’s decisions can accurately predict another judge’s decisions. We find that it cannot: judges exhibit meaningfully distinct decision-making processes. Specifically, when we evaluated each judge-specific model on cases decided by other judges, performance deteriorated substantially, revealing pronounced heterogeneity across judge pairs. We then identified structured “disagreement rules” that characterize when particular judges diverge. For instance, Judge 9 and Judge 4 disagree in 94% of cases involving individuals aged 24 or younger, suggesting systematically different treatment of younger individuals. This indicates that the justice system’s outcomes depend heavily on who the judge is. A triplet consistency experiment, where we matched similar cases for different judges, confirmed these findings: judges were highly self-consistent, getting similar outcomes for similar cases 76% of the time, but only modestly consistent with one another, coming to the same rulings on similar cases only ~55% of the time (close to random guessing), indicating that the justice system’s outcomes depend heavily on who the judge is.

Our results thus reveal a paradox. On one hand, judges do behave algorithmically a lot of the time; their decisions can be captured by small, interpretable formulas that rely on measurable variables. On the other hand, the algorithmic behavior differs from judge to judge, leading to inconsistency and unequal treatment across defendants. Identifying and comparing these implicit rule sets could enable data-driven judicial training, standardized guidelines, and mechanisms to promote interpretability and system-wide fairness.

By understanding when and why judges deviate from interpretable rules, we can better identify errors in data, biases in reasoning, and opportunities for reform. Ideally, with additional data on outcomes, we can focus judges on rules that produce desirable outcomes (Loeffler and Grunwald 2015), while being a comparatively inexpensive type of intervention (Ludwig, Mullainathan, and Rambachan 2024). Ultimately, we claim the central question is not whether algorithms belong in the justice system, but whether humans within it already act like them and whether we can make both better.

Related Works

A growing empirical literature studies judicial decision-making as a form of human judgment under uncertainty, particularly in high-volume and time-constrained settings such as pretrial bail hearings. This work consistently documents substantial variation across judges, even when they are presented with similar cases, raising concerns about consistency, equity, and arbitrariness in legal outcomes.

In the context of bail, several studies compare judicial decisions with algorithmic predictions. Kleinberg et al. (2018) analyze pretrial release decisions in New York City, showing that simple machine learning models can outperform judges in predicting failure to appear while simultaneously reducing incarceration and racial disparities. Other work (Arnold, Dobbie, and Hull 2022) demonstrates that judge-specific preferences contribute meaningfully to variation in bail outcomes, with some judges systematically imposing more restrictive conditions than others, even after controlling for observable case characteristics. These findings highlight the limits of unaided human judgment in repetitive, prediction-driven legal tasks.

These concerns are especially salient in Harris County, Texas, following the *O’Donnell* Consent Decree. Before the decree, misdemeanor bail decisions were governed by a rigid bail schedule that tied release to the ability to pay, resulting in widespread detention of low-income defendants. Post-decree evaluations (Heaton, Mayson, and Stevenson 2017) find that the reforms substantially reduced wealth-based detention, but did not eliminate heterogeneity in magistrate decision-making. In carve-out cases that proceed to individualized bail hearings, outcomes remain sensitive to which magistrate presides, suggesting that discretion continues to play a central role.

Beyond the legal literature, research in cognitive psychology provides insight into why judicial decisions may be well approximated by simple, interpretable rules. A large body of work shows that human experts frequently rely on heuristics—fast and frugal decision rules that trade optimality for cognitive efficiency. Foundational work on availability and representativeness heuristics (Tversky and Kahneman 1974) demonstrates how humans simplify probabilistic reasoning, while other works (Marewski and Gigerenzer 2012) show that in domains such as medical decision-making, heuristic-based models can rival more complex statistical approaches. These findings suggest that judicial reasoning, particularly under time pressure and informational constraints, may naturally take a rule-like or algorithmic form.

Consistent with this perspective, prior studies have found that expert decisions in high-stakes domains can often be captured by compact decision rules. Models of bail decisions in the United Kingdom using fast-and-frugal trees found that judges rely on a small number of cues applied in a consistent order (Dhami and Ayton 2001). Similar results have been observed in other expert domains, including military and medical decision-making, where simple decision trees achieve accuracy comparable to more complex models while remaining interpretable.

Again, we do not ask whether algorithms should replace judges or whether algorithmic predictions outperform those of human decision-makers. Instead, we ask whether judges themselves already behave like algorithms: whether simple, interpretable rules can capture their decisions; whether those rules are consistent over time; and when no reasonable algorithm explains a judge’s behavior. By shifting the focus from human-versus-algorithm comparisons to the internal structure of judicial decision-making, our work connects normative debates about rules and standards with empirical evi-

dence on human judgment, offering a new lens on accountability and consistency in the justice system.

Data and Processing

In Harris County, the Consent Decree gives the public access to certain data concerning misdemeanor pretrial release and detention decisions, including defendant characteristics (such as race, sex, and age), case characteristics (such as charges filed, booking and release dates), which magistrates and judges are assigned to cases, decisions made at bail hearings, and bail outcomes in cases from 2009 to the present. We focused on misdemeanor and felony cases from 2020 to 2025, a time period during which the Consent Decree rules had been adopted and were being used consistently. The dataset contained 531,921 unique cases, out of which we filtered out carve-out cases (cases that were not given a General Order Bond), duplicate cases, and off-docket cases (pre-screened cases where pretrial services had determined the defendant was eligible for release on a personal bond and where the magistrate agrees, granting the personal bond without a bail hearing). Using this data, we combined the crime case details and bond analytics information from the felony dataset and misdemeanor datasets for individual defendants to compute the features in Table 5 in Appendix B.2. These features mostly pertain to that defendant’s criminal history: one for past misdemeanors and one for past felonies (with the ‘_f’ suffix). Felony crime information is included because we discovered that magistrates sometimes grant personal bonds despite pending felony charges, reasoning that the ongoing case would ensure punishment if a reoffense occurred, making it important for modeling their decisions. A full description of the dataset and processing is in Appendix B.

After feature engineering, we excluded cases with any existing holds and deferred adjudications, since magistrates appeared to view these existing constraints as sufficient safeguards when granting personal bonds, as well as cases with null values for any of the features, leading to a total of 22,361 usable cases in our dataset. Out of these, 18,456 cases (82.5%) had personal bond not granted (class 0) and 3,905 cases (17.5%) had personal bond granted (class 1). Figure 1 shows our data processing workflow. After processing, we decided to use 21 of the 24 judges, excluding those who adjudicated too few cases to support reliable analysis. We assigned each included judge an anonymous identifier.

Experiments and Results

We structure our experiments around our four guiding research questions.

When Do Judges Not Follow Any Reasonable Algorithm?

Here, we find decisions that are not consistent with any reasonable algorithm characterizing the decision-making process of the judge, based on the available court data. We call these “unexplainable cases” since these decisions are seemingly incompatible with the judge’s usual way of reasoning. Such cases could arise from a variety of situations: evidence

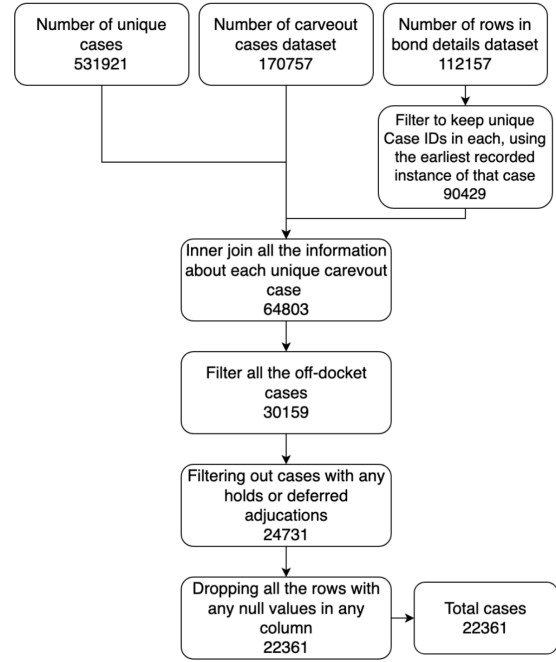


Figure 1: Data Processing Workflow. The flowchart outlines steps for merging and cleaning judicial case records.

not in the database (missing data), incorrect data, personal bias of the judge, corruption, or randomness in decisions.

To find such cases, we used the TreeFARMS algorithm (Xin et al. 2022) to obtain a *Rashomon set* of interpretable trees for each judge. A *Rashomon set* is the collection of models from a given function class (here, simple decision trees) that are the most performant on the dataset. Whereas standard machine learning methods provide one “best” model, we do not, because there can be many models that perform about equally well, and those models differ from each other (Semenova, Rudin, and Parr 2022; Semenova et al. 2023; Fisher, Rudin, and Dominici 2019; Rudin et al. 2024). TreeFARMS is designed to efficiently enumerate and represent the *Rashomon set* of sparse decision trees (Xin et al. 2022) for a given dataset. Using TreeFARMS allows us to study variation across equally good models, examine which features consistently appear as important, and highlight areas of uncertainty or flexibility in decision-making. Details about generating the *Rashomon sets* are in Appendix D. Table 1 shows the results of the *Rashomon set* generation, including the number of trees for each judge and best model accuracies, ranging from 60% to 99%.

We examine some trees drawn from the *Rashomon sets* generated by TreeFARMS. Figure 2 shows an example tree for Judge 17. More examples from the *Rashomon set* of Judge 17 can be found in Appendix F. One notable pattern that we observe for these judges is that they seem to grant personal bonds to individuals with ICE holds. After reviewing the PC forms, we realized that when a higher authority places a hold, the judge typically anticipates that this authority will handle the defendant’s outcome. As a result, the

Judge	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Rashomon Set Size	1188	612	682	40	272	354	76	24	7	16	852	210	1164	976	62	338	117	513	388	28	1088
Best Model Accuracy	78%	70%	73%	66%	73%	79%	66%	84%	85%	99%	61%	83%	78%	69%	98%	83%	65%	78%	83%	85%	80%

Table 1: Summary of Rashomon set size and representative model accuracy across judges.

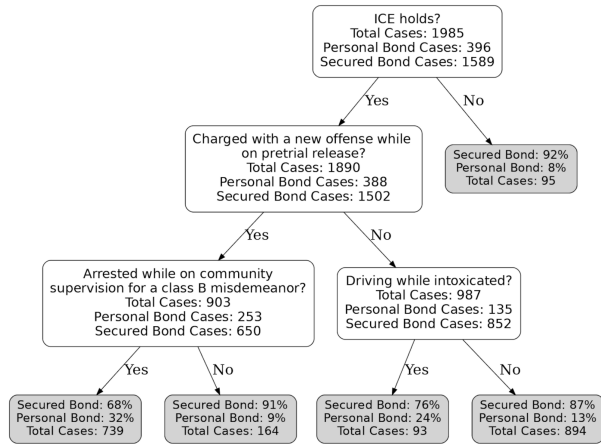


Figure 2: Sample decision tree from the Rashomon set for Judge 17.

judge often releases the defendant on a personal bond. We show trees from other judges later in this manuscript and in Appendix E.

BAIL ORDER

☒ Defendant is ordered RELEASED on a personal bond in the amount of \$2500; OR

☐ The Court ORDERS Defendant committed to the custody of the Sheriff of Harris County, Texas,

☐ bail was denied after a hearing; OR

☒ until Defendant posts the required bond or until further order of the Court, AND:

☒ Defendant must pay \$2500 secured bail to be released and I find that Defendant can afford the secured bail amount; OR

☐ Defendant must pay \$ secured bail to be released. Defendant does not have the ability to pay the amount required, but I considered alternative conditions of release, and by clear and convincing evidence, there was no less-restrictive condition or combination of conditions that could reasonably prevent flight from prosecution and/or reasonably protect the safety of the community; OR

☐ Defendant must pay \$ secured bail to be released. I could not determine indigency and ability to pay because Defendant did not swear to and sign a financial affidavit, and Defendant declined to provide evidence related to indigency and ability to pay during the hearing. I considered alternative conditions of release, and by clear and convincing evidence, there was no less-restrictive condition or combination of conditions that could reasonably prevent flight from prosecution and/or reasonably protect the safety of the community.

Reasons for the bail decision and evidence relied upon: on bond for asslt fv PRBD 06-06-06 same CW, share a child dating relationship, alleged D pushed CW to the ground and kicked her in the head, FTA BR 07-07-07 got another PRBD AACOB no threatening, harassing or assaultive contact w CW // NIVs asslt fv and DWI 2nd, W called police about argument turning physical in truck, RO arrived CW hanging out of truck, CW said D mad bc thought she was w another guy last night, he was drinking, said D bit her and was pushing her, officer observed bit mark, D had signs of intoxication, 12 pack of Modelo beer on passenger side floorboard // lethality factors: intimate partners, substance abuse, power and control // D continues to engage in conduct that results in his arrest while failing to appear in court and while violating conditions of his bond put into place to reasonably ensure the safety of this particular complainant // CW did not request MOEP, MOEP issued based on lethality factors and D on bond for assaulting same CW

I certify that the procedures and findings required by Local Rule 9 were provided and that I have explained the reasons and evidence relied on for my decision to Defendant and/or counsel. ☒ orally, ☐ in writing, ☐ or both.

Figure 3: Sample PC form for a case where there seems to be a discrepancy in marking the correct bond type.

Reasons for the bail decision and evidence relied upon: defendant says was in custody elsewhere and that is why missed court, but he knows what day he supposed to be in court and he will be there to resolve these cases

Figure 4: Sample PC form for case where judge had information outside of our dataset influencing their decision.

Reasons for the bail decision and evidence relied upon: has a pending VPO in Richford county 06-06-06 TTFV 07-07-07 Richford // 2yrs TDC agg asslt fv 2001 and asslt fv impeding was dismissed as part of plea agt, deadly conduct 1925 // cw is defendant's blind cousin he was hitting cw in the face and with a chain other family members say he always does this when he gets drunk // then was calling the officer the n-word repeatedly and giving the only good place for the black officer is swinging from a tree and that he is going to find him and kill him // MOEP requested that I can not issue

Figure 5: Sample PC form for a case where judge had information about pending charges outside of our dataset, meaning we had incorrect features in our dataset.

Reasons for the bail decision and evidence relied upon: Motions for Bond Conditions and for MOEP / HC: none / Out of county DWI / no prior violent offenses / Sgt pulls up on car where Def is being argumentative with the CW Female. Def says that female will not let him see the baby, there is a baby in the back cr seat. CW Jane Doe says that she is afraid of Def because Def threatens her

and she is afraid that he will follow thru with the threats. Sgt was shown threatening text messages. Def threatened to f-up CW if she did not bring the child to HEB. She went to HEB and Def immediately became belligerent and aggressive. When CW tried to step out of the car, Def slammed the door on her ankle, causing her pain, swelling and leaving a mark. Grant MOEP and order conditions.

Reasons for the bail decision and evidence relied upon: mtms bond conditions / hc: none / def assaulted his wife and 3 month old child, cw was cooking, def came home and said he did not like what she was cooking, argument, cw blocked doorway so he could not leave, def punched her in the face, def hit the child several times as well, she told him to leave, def punched her some more,

def dragged her and dropped and she hit her head, def punched her again when she stood up, wife had multiple injuries, child had a mark on her cheek, def is not indigent, def is a risk at a minimum to his wife. There is clear and convincing evidence that the arrestee does not have the ability to pay the amount required, but alternative conditions of release were considered, no less-restrictive condition or combination of conditions could reasonably assure the safety of the community or reasonably protect against flight from prosecution, and imposition of unaffordable money bail is necessary to reasonably assure the safety of the community or to rea

Figure 6: Sample PC forms for cases in which crimes share similar features but result in different outcomes due to differences in crime intensity, a feature we do not have.

Next, we used our Rashomon sets to generate subsets of each judge's cases that could not be explained (i.e., predicted correctly) by any model in the Rashomon set. These cases, about 18% of all cases, are "unexplainable." We investigated a sample of these cases (100) manually and found that all of them were unexplainable due to noise.

We read the judges' justifications and discovered that such cases fell into three main categories: (a) Dataset errors (form errors); (b) Missing information about other crimes (outside of Harris County) that are not in our database; (c) Details about the case that are not in the database but are available to the judge.

Figure 3 shows a form error. Our dataset indicates that a personal bond was granted; however, the judge's comments suggested against granting a personal bond. When checking the form, we discovered that multiple boxes had been checked for the type of bail order, which must be an error since one cannot receive both a personal and a secured bond in the same hearing. The second category includes cases where the judge has information outside of Harris County, which is logistically difficult for us to obtain. Figure 4 shows

an example where the judge knew that the defendant had been in custody in another jurisdiction, which influenced their decision to grant a personal bond; this information was not available to us in the court data. The last category includes cases where the judge makes decisions using details not in our structured data (usually information presented during the hearing but not previously available). Figure 6 shows a pair of such cases, where both crimes are Class A misdemeanors of family assault; however, the judge states that they granted a personal bond in the first case and not in the second due to the intensity of the crimes.

All three categories lead to flaws in the data that will not help us understand judges. Ideally, we use data for building models that is identical to the data the judges use. Form errors simply provide incorrect information. Missing information from other counties results in incorrect features in our dataset. E.g., if someone commits a crime in a different county, our criminal history features would not record it, creating an error in our data.

The final category is case-specific missing information, such as the severity of the crime, which is not recorded as part of the penal code and crime type (e.g., “family assault”); this could be remedied if it were consistently mentioned, for instance. Including these three classes of cases would require our models to predict decisions using an incomplete representation of the judge’s information. This may cause models to attribute the effects of these unobserved factors to the available features, leading to biased estimates of feature importance and poorer generalization. Excluding these noisy cases allows models to be trained on observations where the recorded features mainly reflect the available information during a judicial decision and avoids diluting results with noise.

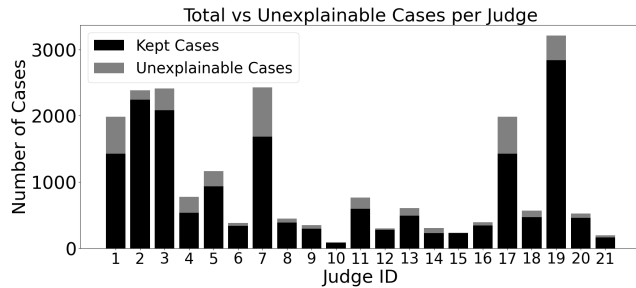


Figure 7: Per-judge distribution of unexplainable vs kept cases.

The number of unexplainable cases compared to total cases is shown in Figure 7, and the class distribution before and after filtering is illustrated in Figure 21 in Appendix G. In what follows, we will remove all the unexplainable cases and only use the remaining cases.

Can Each Judge Be Described by an Algorithm?

If a judge follows a systematic method – an algorithm – to determine whether to grant a personal bond, we should be able to figure out what that algorithm is. Examining this

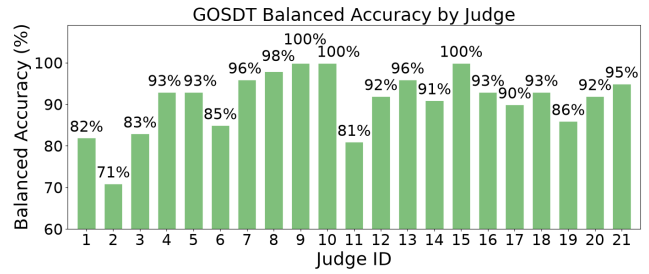


Figure 8: Balanced accuracy for the GOSDT models trained for each judge.

question allows us to evaluate judicial decision making, including whether the rules judges follow in practice are desirable. Again, we are *not* arguing for replacing judges with algorithms, but rather seeking to identify when and whether judges already behave like algorithms.

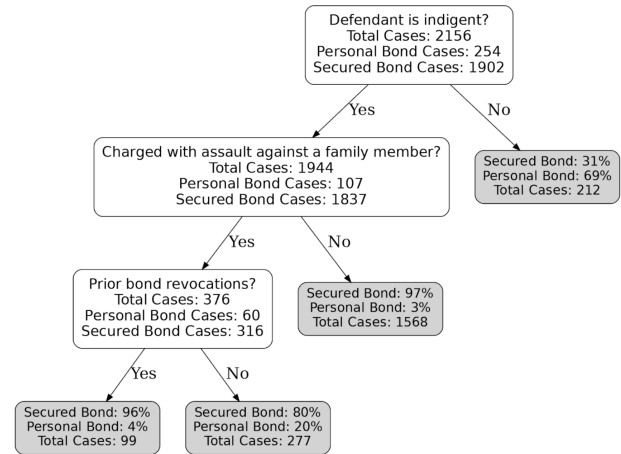


Figure 9: Decision tree for Judge 1 trained using GOSDT.

To answer this question, we trained representative models using GOSDT (Lin et al. 2020) on the filtered subset of cases per judge, producing globally optimal sparse decision trees. As discussed, decision trees are a powerful, non-parametric function class whose models are understandable. Each magistrate’s model was fitted with a tree depth budget of 5, regularization set to 0.001 (a small value, to allow for higher accuracy), and class balancing enabled. This yielded concise, interpretable rules that mimic each judge’s decision style. We report the balanced accuracy in Figure 8 and show visualizations of some of the resulting decision trees with per-node class counts to interpret the logic of each model.

We discovered that most of the judges could indeed be explained reasonably well by simple algorithms. Figure 25 in Appendix J shows the model for Judge 4 with an accuracy of 91.79% and a balanced accuracy of 92.88%. That is, about 92% of Judge 4’s cases agree with this decision tree, and it is accurate in explaining decisions for both personal bonds and secured bonds. For 17 of the 21 judges, the models capture at least 85% of the judge’s decisions. Another example

of a good model is presented in Figure 9 for Judge 1, with an accuracy of 84.37% and a balanced accuracy of 82.44%. Figure 26 in Appendix J shows a model that explains Judge 16 well, with an accuracy of 92.46% and a balanced accuracy of 93.06%.

There are several judges whose decision making we cannot explain as well. Figure 27 in Appendix K shows an example of a tree for Judge 2 that has clearly been overfit (with fewer than 20 cases per leaf) and yet has an accuracy of 77.48% and a balanced accuracy of 75.41%. Even after overfitting to the data, 23% of the cases have no simple rule characterizing them.

For judges that can be explained, we next examine which factors they seem to consider when making bail decisions.

Do Judges Consider the Same Variables?

In this section, we will show that judges generally consider similar types of variables but differ in the importance of specific variables.

For each judge, we used a bootstrapped feature selection algorithm. For each bootstrap, we balanced the data by over-sampling the minority class. Then, we used the Threshold Guessing Binarizer (which uses a boosted tree classifier) described in McTavish et al. (2022), using 50 estimators and a max depth of 3, to transform the variables into binary indicators. Features that did not appear from the Binarizer are unimportant to boosted trees and were removed. Then, we used a boosted tree classifier again on the remaining features. We used this classifier to obtain variable importance scores for each original variable x_j , whose importance is denoted by I_j . We assigned $I_j = 0$ for all unselected features.

Across bootstrap resamples, the selection frequency of feature j is $F_j = \mathbb{P}(I_j > 0)$, the proportion of bootstraps in which feature j is used. The final feature score is $S_j = I_j \cdot F_j$, and we select the top 50 features by S_j in descending order. From these 50 features, we computed the following average variable importance metrics over 100 bootstraps: permutation importance, conditional model reliance (CMR), and leave one covariate out (LOCO). We also computed the accuracy of a model that uses these top 50 features, and we repeated this for each bootstrap and averaged the result over bootstraps. The result is in Figure 10.

As we have reduced the noise by filtering out unexplainable cases, we intentionally overfitted the models to the data to ensure that the variables are meaningfully important to the data we have. We see in Figure 10 that the average accuracy for most judges is extremely high. Reviewing the variables among all judges in Table 2, we see that age is *extremely* important, being important for every single judge. Meanwhile, most judges use warrant count and “rule” variables, which we define as flags for specific crimes that are algorithmic in nature for bond trials, indicating that many judges are using them correctly. Furthermore, we can see that race is only important to some judges, allowing us to uncover potential biases that different judges may have. Furthermore, the collection of variables is not consistent among judges, suggesting that judges generally have quite different values when making pretrial decisions. For example, when we com-

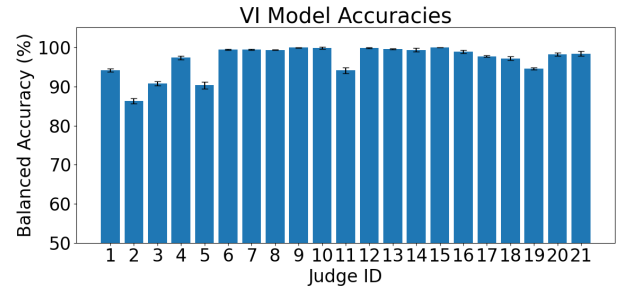


Figure 10: Model accuracy, built from top 50 features, averaged over bootstraps, for each judge. Standard deviations are also shown.

pare the top 10 variables for Judges 4 and 6 in Table 3, we see a trend where they share many common variables, such as age, warrant count, and “rule” variables, but also have unique variables that each weighs heavily. Judge 6 weighs family assault and criminal history more heavily than Judge 4. In terms of demographics, both judges consider age and citizenship status, but Judge 6 appears to weigh gender and homelessness much more than Judge 4, showing how we can use this framework to see the differences in what different judges value.

Feature	Appear.	Mean rank	Prop.	Score
Age	21	3.238	1.000	0.309
Warrant Count	19	5.263	0.905	0.172
Charged with a new offense while on pre-trial release? (rule)	15	4.800	0.714	0.149
Arrested after a bond forfeiture or revocation? (rule)	13	7.359	0.619	0.084
Arrested while on supervision for a class B misdemeanor? (rule)	11	8.136	0.524	0.064
First warrant is a CAPIAS warrant?	10	8.667	0.476	0.055
African American?	12	11.444	0.571	0.050
Violation of bond conditions for penal code 25.07 (rule)	9	9.500	0.429	0.045
Family Assault?	9	10.519	0.429	0.041
How many past felonies?	9	11.037	0.429	0.039

Table 2: Variable importance summary across judges. Features are ordered by score, defined as proportion/mean_rank, where *mean_rank* is the average rank when the feature appears in a judge’s model and *proportion* is the fraction of judges for which the feature appears in the top 15.

Could All Judges Be Using Similar Algorithms? Are Judges Consistent With Themselves and Each Other?

In order to ensure fairness in our justice system, we want judges to make decisions consistently. That is, presented with the same information, we want them to make the same decision. We also want different judges to make the same

Judge 4				Judge 6			
Feature	Rank	CMR	LOCO	Feature	Rank	CMR	LOCO
Age	1	1	1.0	Age	2.0	2	1.0
Arrested after a bond forfeiture or revocation? (rule)	4	3	3.0	Female?	1.0	1	10.0
Criminal Trespassing?	3	4	4.0	Harrasement with previous conviction	4.0	4	4.0
Assault with Bodily Injury?	5	5	2.0	Family Assault?	7.0	5	12.5
ICE hold	2	2	8.5	Unknown Citizenship	13.0	14	3.0
Unknown Citizenship	6	6	5.0	Charged with a new offense while on pretrial release? (rule)	17.0	10	5.0
Warrant Count	7	7	7.0	Warrant Count	15.0	19	2.0
Charged with a new offense while on pretrial release? (rule)	12	8	10.0	Penal Code 28.03	11.0	9	17.5
Violation of bond conditions for penal code 25.07 (rule)	10	12	15.0	Homeless?	3.0	3	33.0
First warrant is a CAPIAS warrant?	14	13	11.0	How many past felonies	9.0	17	14.0

Table 3: Variable importance ranks for Judge 4 (left) and Judge 16 (right) for the top 10 variables. The variables are ordered by mean ranking over the three importance metrics for each judge, with rank referring to permutation importance.

decisions when presented with similar information; i.e., we want all judges to follow the same algorithms. In this section, we hope to analyze the consistency of a judge’s own decisions and those of others, as well as provide a way to give feedback to current judges and judges in the future. To do this, we performed three different sets of experiments.

Are Judges Consistent with Each Other? Pairwise Loss Distribution Analysis. We carried out a cross-judge evaluation to examine consistency in decision rules across judges. For each judge, we took the full Rashomon set of near-optimal decision trees and used them as source models. Every model was evaluated not only on the judge’s own cases but also on the cases belonging to every other judge. To make this feasible, we standardized the input space by re-constructing the engineered features for all judges and aligning them by case ID. Any features expected by a source model but missing in a target judge’s dataset were added as zero columns, ensuring comparability across evaluations. Losses were computed separately for the source judge’s data and for each target judge’s data. By repeating this for all models in every Rashomon set, we obtained a full distribution of losses rather than a single point estimate. These distributions were visualized with violin plots.

We found that most of the judges are not consistent with one another. For example, Figure 11 shows the violin plot of the loss distribution for all the models within the Rashomon set of Judge 16 when evaluated on their own cases (source cases) versus those of Judges 1 and 8 (target judges). We see that the models perform consistently well on the source data, with the error rates clustering tightly around 6%, which indicates coherent decision logic. However, when the same models are applied to the target judges, the losses see a sharp rise, with larger ranges and more errors for the target judges. Although the model found for Judge 16 captures multiple near-optimal explanations for their own decisions, these explanations do not apply to the decisions made by Judges 1 and 8. This contrast suggests that both judges behave differently from Judge 16 in their decision-making processes.

However, in some judge pairs, the difference between the source and target distributions is less pronounced, as shown in Figure 28 (left) in Appendix L. Here, the Rashomon models of Judge 2 maintain comparable performance when evaluated on Judge 3, with overlapping distributions and similar median error rates of 0.24 to 0.27. While the variance increases slightly when predicting on the target judge’s cases,

the overall shape of the distributions suggests partial alignment in decision behavior between the two judges. This implies that both judges may rely on the same features and follow similar reasoning patterns when deciding cases. Figure 28 (left) also shows that Judge 2 may be internally inconsistent since their loss distribution seems to be quite spread-out. We also found several examples with partial overlap. The violin plots in Figure 28 (right) show that while the Rashomon models of Judge 6 perform reasonably well on their own decisions, their performance varies widely when applied to Judge 15. The source distribution is reasonably compact, with a median error around 0.20. In contrast, the target distribution is widest near both the lowest and highest ends of the loss scale, suggesting a bimodal pattern of generalization. Some models from the source set transfer well, achieving low errors on the target judge, while others perform poorly, reaching error rates above 0.50. This pattern could imply that the two judges share overlapping logic on certain subsets of cases but may diverge on others.

Loss Distributions: Judge 16 (source) vs Judges 1 and 8 (targets)

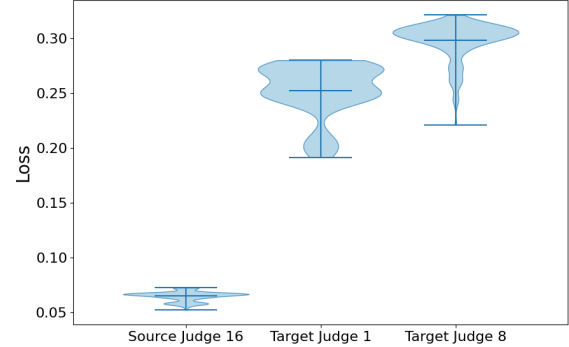


Figure 11: Rashomon model loss distributions on Judge 16 source cases. Compares the source-judge model (S16) with target-judge models (T1 and T8).

When Do Judges Disagree? Rule Mining Analysis. We then conducted rule mining to extract interpretable conditions under which specific pairs of judges’ decision rules diverged. For each ordered judge pair, we used majority-vote predictions from both Rashomon sets on the source judge’s cases to define a binary disagreement variable, which would be true if the labels did not match and false otherwise.

In rule mining, **support** measures how frequently an item-set appears in the dataset, **confidence** indicates how often the rule’s consequent occurs given its antecedent, and **lift** quantifies how much more likely the consequent is to occur with the antecedent than it would be by chance. The

Judge A	Judge B	Disagreement Rules	Supp.	Conf.	Lift	Cases
Judge 4	Judge 7	Age $\leq$ 24 AND assault on a family member AND $\leq$ 1 prior warrant	0.193	0.935	1.70	305/326
Judge 1	Judge 17	Assault on a family member AND $\leq$ 1 prior felony AND new offense on supervision	0.229	0.852	2.09	420/493
Judge 9	Judge 14	$\leq$ 1 past crime	0.623	0.750	1.43	138/184
Judge 2	Judge 18	Indigent AND first arrest due to non-appearance AND $\leq$ 1 prior warrant	0.195	0.769	1.81	337/438
Judge 7	Judge 9	first warrant issued for crime commitment	0.356	0.757	1.15	69/105

Table 4: Five sample rules identified by FP-Growth that characterize situations where pairs of judges disagree.

antecedent space is defined as the union of the engineered binary features used by both judges. We mined candidate rules of length one to three under minimum support (0.02 – meaning 2% of the data must obey the left side of the rule) and minimum confidence (0.60 – meaning at least 60% of the right side of the rule obeys it) thresholds, with the top 50 rules per length.

We used FP-Growth (Han, Pei, and Yin 2000) to generate our set of rules for each judge pair. Each rule was then used on the underlying cases to compute canonical counts of covered cases, disagreements, and agreements, from which support, confidence, and lift were recomputed relative to the global disagreement rate. This procedure, applied in both directions for all judge pairs, yielded a set of concise, understandable rules highlighting the specific feature combinations most predictive of inter-judge disagreement.

We find that most judges with distinct loss distributions exhibit strong, interpretable rules that explain their disagreements. For example, the rule “age $\leq$ 24 $\rightarrow$ Judges 9 and 4 disagree” is highly predictive (support = 0.18, confidence = 0.94, lift = 1.67), indicating that disagreements between Judges 9 and 4 occur in nearly all cases involving defendants aged 24 or younger. Similarly, the rule “charged with family assault AND charged with new offense on pretrial release $\rightarrow$ Judges 21 and 17 disagree” identifies a substantial subset of cases: 24% satisfy both conditions, and the judges disagree in 85% of them. These patterns suggest that outcomes in such cases are driven more by individualized judicial discretion rather than by standardized risk cues. These rules can identify specific case-qualities where judges most often disagree, providing an avenue to under-

stand when judges disagree and potentially reduce inconsistencies in their decision-making by aligning judge’s decisions in these scenarios.

Additional examples are in Table 4.

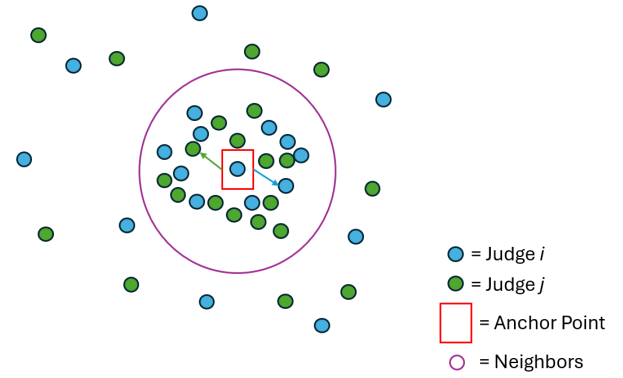


Figure 12: Triplet Example. This is an example of a triplet, where the blue points are from Judge i and the green points are from Judge j . Given an anchor point from Judge i , a triplet is then made from choosing a random nearest neighbor from Judge i and Judge j . In this example, any point within the purple circle is available to be chosen as a random nearest neighbor for the triplet.

Are Judges More Consistent with Their Own Decisions Than with Others? Triplet Analysis.

Even if two different judges follow different algorithms, it is still important to see whether they make similar decisions. To do this, we performed a triplet analysis that compares two judges. For each pair of judges, there is an anchor judge and a target judge. A triplet is defined by an anchor case from the anchor judge, a nearest neighbor case from the anchor judge, and a case from the target judge that is similar to the anchor case. We define similar cases to be cases with a low normalized weighted Euclidean distance, where highly important variables like offense type, rules, and age are weighted more heavily than other variables. Figure 12 shows a visualization of how these triplets are formed. In this triplet analysis, we randomly sampled ~ 1000 triplets for each pair of anchor and target judges. Analyzing these triplets, we consider the following questions: *When two judges disagree, does each judge predict themselves better than the other?*, *When two cases presented to the same judge are similar, do they make the same decision?*, and *Does the choice of judge affect the decision of a case?* Let us answer these. We use the same triplets in our entire analysis so that we can directly compare our metrics.

When two judges disagree, does each judge predict themselves better than the other?

To evaluate this, we calculate the percentage of triplets where Judge i makes the same decision as the anchor point, meaning Judge i predicts themselves correctly, while Judge j disagrees. We define this metric as the win rate. If a judge predicts themselves better

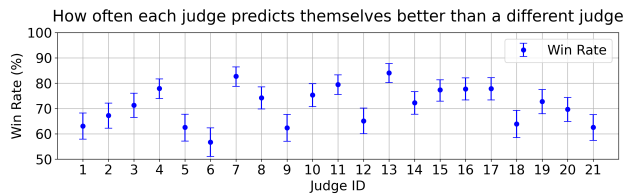


Figure 13: Average win rates of judges. All win rates exceed 0.5, indicating that judges predict their own outcomes better than those of other judges in the majority of cases.

than other judges do, they have a high win rate. We see in Figure 13 that most judges have a win rate between 60% and 80%, meaning that each judge is a better predictor of themselves than other judges most of the time. While this is good, we also see that there can be improvement, as the win rates cap out at around 85%.

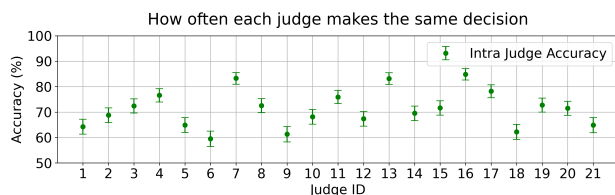


Figure 14: Judge’s consistency in decision-making. The average percentage of triplets where a judge predicts the same outcome as the anchor case. All judges are above 50%, indicating performance better than random guessing, with most around 60–70% and some near 80%.

When two cases presented to the same judge are similar, do they make the same decision? For a fair justice system, we expect a judge to always make the same decision when given the same case. To evaluate whether this happens or not, we find the percentage of triplets (using just the anchor and the other point from the same judge) where the anchor judge makes the same decision. As we see in Figure 14, most judges have a percentage between 60% and 80%, indicating that while they are generally more consistent than random, there is still a large gap between judges always being consistent with themselves. This is an important evaluation to ensure that judges are maintaining consistency.

Does the choice of judge affect the decision of the same case? While we cannot test giving two different judges the same case, we can use our triplets to approximate this. To evaluate whether the choice of judge matters, we find the percentage of triplets (using just the anchor from Judge i and the point from Judge j) where Judge i makes the same decision as Judge j . A higher percentage means the choice of judge does not matter, while a low percentage near 50% suggests that it does. As we see in Figure 15, the percentage is about 50%, maxing out at $\sim 60\%$, which is the minimum percentage for both the win rate and how consistent each judge is with themselves. This means that the choice of

judge is *very* important in deciding whether a defendant will receive a personal bond or not.

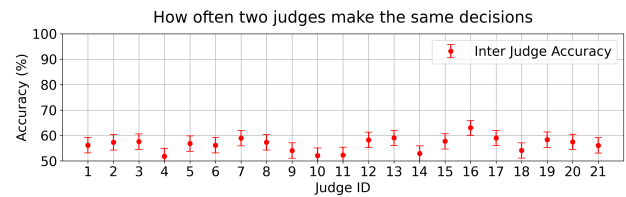


Figure 15: Consistency between judges. The average percentage of triplets where a judge predicts the same outcome as the second judge in the triplet. We see that all judges are slightly better than random guessing. This shows that agreement between judges is somewhat random.

Discussion

In our justice system, we expect judges to make decisions consistently. Thus, we expect judges to follow *rules*: decision patterns using algorithmic logic. However, some cases require the application of *standards*, i.e., contextual understanding and moral reasoning that computational models cannot replicate. If judges do follow rules, we can potentially improve them through better predictive modeling; improving rules is comparatively inexpensive and quite feasible (Ludwig, Mullainathan, and Rambachan 2024), and would make judicial decisions more accurate (DeMichele et al. 2021) and transparent. If judges do not follow algorithms, it is important to know why, because arbitrariness in judicial decision-making is undesirable. We have provided and used a framework that evaluates how often judges follow algorithms and how and whether their algorithms differ. We saw that judges often follow algorithms and that different judges follow different algorithms, which produce different outcomes in similar cases. This means judges could potentially be made more consistent and accurate. The framework also includes a variable importance analysis that can uncover whether judges may heavily consider features like race or other demographic characteristics to make decisions. We also found many cases where missing or erroneous data required the use of standards rather than rules. Better data collection could then lead to more consistent, rule-based decisions. Given high quality recidivism data, it is possible to extend the framework by evaluating judge’s algorithms to determine whether judges accurately predict long-term outcomes.

Instead of viewing algorithms as replacements, they should be seen as instruments to inform and enhance human judgment. No one can accurately estimate risks from sufficiently large and complex data in their head, but ensuring human oversight remains central, especially in high-stakes cases that could determine an individual’s freedom. By making judges’ internal rules, i.e., algorithms, as well as their reliance on standards, more explicit and accountable, we can move towards a judicial system that is both transparent and self-improving.

Acknowledgements

We thank Arvinhd Manian for his contributions to the early stages of this work.

References

- Arnold, D.; Dobbie, W.; and Hull, P. 2022. Measuring Racial Discrimination in Bail Decisions. *American Economic Review*, 112(9): 2992–3038.
- Crotofof, R.; Kaminski, M. E.; Price, W.; and Nicholson, I. 2023. Humans in the Loop. *Vand. L. Rev.*, 76: 429.
- Deeks, A. S. 2019. The Judicial Demand for Explainable Artificial Intelligence. *Virginia Public Law and Legal Theory Research Paper No. 2019-51*.
- DeMichele, M.; Comfort, M.; Barrick, K.; and Baumgartner, P. 2021. The Intuitive-Override Model: Nudging Judges Toward Pretrial Risk Assessment Instruments. *Federal Probation*, 85(2).
- Dhami, M. K.; and Ayton, P. 2001. Bailing and jailing the fast and frugal way. *Journal of Behavioral Decision Making*, 14(2): 141–168.
- Eaglin, J. M. 2023. Racializing Algorithms. *California Law Review*, 111: 753.
- Fagan, F.; and Levmore, S. 2019. The impact of artificial intelligence on rules, standards, and judicial discretion. *Southern California Law Review*, 93: 1.
- Fisher, A.; Rudin, C.; and Dominici, F. 2019. All Models are Wrong, but Many are Useful: Learning a Variable’s Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, 20(177): 1–81.
- Garrett, B. L. 2025. Artificial Intelligence and Procedural Due Process. *Duke Law School Public Law & Legal Theory Series No. 2025-15*.
- Garrett, B. L.; and Rudin, C. 2024a. The Right to a Glass Box: Rethinking the Use of Artificial Intelligence in Criminal Justice. *Cornell Law Review*, 109: 561–627.
- Garrett, B. L.; and Rudin, C. 2024b. Testing AI. *Duke Law School Public Law & Legal Theory Series No. 2024-62*.
- Han, J.; Pei, J.; and Yin, Y. 2000. Mining frequent patterns without candidate generation. *SIGMOD Rec.*, 29(2): 1–12.
- Harcourt, B. E. 2015. Risk as a Proxy for Race: The Dangers of Risk Assessment. *Fed. Sent’g Rep.*, 27(4): 237–243.
- Heaton, P.; Mayson, S.; and Stevenson, M. 2017. The Downstream Consequences of Misdemeanor Pretrial Detention. *Stanford Law Review*, 69: 711.
- Kaplow, L. 1992. Rules Versus Standards: An Economic Analysis. *Duke Law Journal*, 42: 557–629.
- Kleinberg, J.; Lakkaraju, H.; Leskovec, J.; Ludwig, J.; and Mullainathan, S. 2018. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1): 237–293.
- Lin, J.; Zhong, C.; Hu, D.; Rudin, C.; and Seltzer, M. 2020. Generalized and Scalable Optimal Sparse Decision Trees. In *International Conference on Machine Learning (ICML)*.
- Loeffler, C. E.; and Grunwald, B. 2015. Processed as an Adult: A Regression Discontinuity Estimate of the Crime Effects of Charging Nontransfer Juveniles as Adults. *Journal of Research in Crime and Delinquency*, 52(6): 890–922.
- Ludwig, J.; Mullainathan, S.; and Rambachan, A. 2024. The Unreasonable Effectiveness of Algorithms. *AEA Papers and Proceedings*, 114: 623–27.
- Marewski, J. N.; and Gigerenzer, G. 2012. Heuristic decision making in medicine. *Dialogues in Clinical Neuroscience*, 14(1): 77–89. PMID: 22577307.
- McTavish, H.; Zhong, C.; Achermann, R.; Karimalis, I.; Chen, J.; Rudin, C.; and Seltzer, M. 2022. Fast sparse decision tree optimization via reference ensembles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 9604–9613.
- Pepper, J.; Petrie, C.; and Sullivan, S. 2010. Measurement Error in Criminal Justice Data. *Handbook of Quantitative Criminology 353* (A. Piquero & D. Weisburd eds., 2010).
- Rachlinski, J. J.; and Wistrich, A. J. 2017. Judging the Judiciary by the Numbers: Empirical Research on Judges. *Annual Review of Law and Social Science*, 13: 203–229.
- Re, R. M.; and Solow-Niederman, A. 2019. Developing Artificially Intelligent Justice. *Stanford Technology Law Review*.
- Rudin, C.; Zhong, C.; Semenova, L.; Seltzer, M.; Parr, R.; Liu, J.; Katta, S.; Donnelly, J.; Chen, H.; and Boner, Z. 2024. Amazing Things Come From Having Many Good Models. In *Proceedings of the International Conference on Machine Learning (ICML)*.
- Schlag, P. J. 1985. Rules and Standards. *UCLA Law Journal*, 33: 379.
- Semenova, L.; Chen, H.; Parr, R.; and Rudin, C. 2023. A Path to Simpler Models Starts With Noise. In *Proceedings of Neural Information Processing Systems (NeurIPS)*.
- Semenova, L.; Rudin, C.; and Parr, R. 2022. On the Existence of Simpler Machine Learning Models. In *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*.
- Tversky, A.; and Kahneman, D. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157): 1124–1131.
- Xin, R.; Zhong, C.; Chen, Z.; Takagi, T.; Seltzer, M.; and Rudin, C. 2022. Exploring the Whole Rashomon Set of Sparse Decision Trees. In *Neural Information Processing Systems (NeurIPS)*.