5-2024

Computationally Assessing Suspicion

Wesley M. Oliver

University of Cincinnati Law Review

Volume 92 | Issue 4 Article 13

May 2024

Computationally Assessing Suspicion

Wesley M. Oliver
Thomas R. Kline School of Law of Duquesne University

Morgan A. Gray
University of Pittsburgh School of Computing and Information

Jaromir Savelka
Carnegie Mellon University School of Computer Science

Kevin D. Ashley
University of Pittsburgh School of Law; University of Pittsburgh School of Computing and Information

COMPUTATIONALLY ASSESSING SUSPICION*

*Wesley M. Oliver, Morgan A. Gray, Jaromir Savelka & Kevin D. Ashley***

CONTENTS

* The authors wish to thank Mike Livermore, Aileen Nielsen, Andrea Roth, Eric Talley, and participants at the Second Annual Fordham, University of Virginia, and ETH Zurich Data Science and Law Conference for very helpful comments on a previous draft of this Article. We also recognize the excellent research assistance of Rachel Schade.

** Wesley M. Oliver is Professor of Law, Thomas R. Kline School of Law of Duquesne University. Morgan A. Gray is a Ph.D. Candidate in Intelligent Systems, University of Pittsburgh School of Computing and Information. Jaromir Savelka is a Computer Science Associate Research Fellow, Carnegie Mellon University School of Computer Science. Kevin D. Ashley is Professor of Law and Intelligent Systems, University of Pittsburgh School of Law and School of Computing and Information.

INTRODUCTION

As ChatGPT and predictions of a new artificial intelligence-driven society continue to make headlines, the questions of how, and whether, to incorporate artificial intelligence ("AI") into the law are newly urgent. In the context of criminal justice, scholars have already argued that AI has entrenched racial and other biases, such as in algorithmic risk assessments and facial recognition technology.[1] On the other hand, a small number of scholars—including some of the very scholars who have worried about the replication of systemic bias[2]—have suggested that AI might be able to combat bias[3] or improve fealty to the law.[4]

This Article contributes to that discussion by exploring how AI might reduce illegal detentions in drug interdiction stops. Officers tasked with looking for interstate drug traffickers stop motorists on the highway for ordinary traffic infractions and look for "reasonable suspicion" to detain

1. *See, e.g.*, Jessica M. Eaglin, *Racializing Algorithms*, 111 CALIF. L. REV. 753 (2023) (asserting algorithm risk assessment tools reinforce long-existing racial assumptions); Sandra G. Mayson, *Bias in, Bias out*, 128 YALE L.J. 2218 (2019) (contending that predictions about future criminality based on past information necessarily includes racial biases that have long been a part of the criminal justice system); Ngozi Okidegbe, *Discredited Data*, 107 CORNELL L. REV. 2007, 2010-11 (2022) ("[T]he increased use of pretrial algorithm has tended to reproduce racial and socioeconomic inequities."); Andrew Guthrie Ferguson, *Facial Recognition and the Fourth Amendment*, 105 MINN. L. REV. 1105, 1167-91 (2021) (contending that error rates in facial recognition technology, which is higher when a machine is trying to match images for minorities, raise ethical and constitutional concerns).

2. *See* Mayson, *supra* note 1, at 2297 (concluding that the major problem with algorithmic predictions is the prediction of the future from a criminal justice system that had not been racially even-handed, but noting that algorithms are often preferable to human judgment as "reject[ing] algorithms in favor of subjective prediction is to discard the clear mirror for a cloudy one"); Okidegbe, *supra* note 1, at 2061-64 (proposing methods for developing data sets for risk-assessment calculations produced by "community knowledge sources" that do not suffer from the same biases as the information currently employed to develop bail algorithms).

3. *See, e.g.*, Bennett Capers, *Policing, Technology, and Doctrinal Assists*, 69 FLA. L. REV. 723, 755-58 (2017) (arguing that facial recognition technology and big data mining can increase accuracy in identifying those possessing weapons, reducing the often racially-biased determinations made by police officers without these tools); I. Bennett Capers, *Race, Policing, and Technology*, 95 N.C. L. REV. 1241, 1275-76 (2017) (arguing for greater technological surveillance of everyone using cameras and terahertz scanners to "deracialize" the police by improving their ability to assess criminal and innocent behavior by persons of all races); Ric Simmons, *Quantifying Criminal Procedure: How to Unlock the Potential of Big Data in Our Criminal Justice System*, 2016 MICH. ST. L. REV. 947, 950-51 (describing reasonable suspicion and probable cause as "imprecise and subjective" legal standards that could be improved using big data algorithms which "can be structured so that they are truly race neutral and take into account individualized conduct").

4. *See, e.g.*, Andrew Guthrie Ferguson & Richard A. Leo, *The Miranda App: Metaphor and Machine*, 97 B.U. L. REV. 935, 949 (2017) (advocating for a presentation of *Miranda* warning with video, text, and question and answer prior to interrogation, rather than by individual officer, to better convey the information by "remov[ing] the investigative bias that police interrogators introduce into the pre-interrogation process"); Andrew Guthrie Ferguson, *Big Data Prosecution and Brady*, 67 UCLA L. REV. 180, 244 (2020) (proposing the use of predictive analytics to identify exculpatory material that the prosecution must disclose to the defense).

a car until a trained dog can sniff for the presence of drugs.[5] Traffic encounters are dangerous for police and motorists alike.[6] Prolonging a stop to investigate drug trafficking extends these dangerous encounters to a point when the tension from the initial stop has been heightened for both the officers and motorists involved.[7] If the officer incorrectly believes there is justification to continue the investigation, this enhanced risk will have been for nothing. A look at reported decisions in these cases reveals that, not infrequently, courts find that officers have incorrectly evaluated reasonable suspicion and therefore illegally extended a traffic stop for further investigation. Fewer unlawful detentions decrease the amount of evidence lost to suppression hearings claiming that the drugs were discovered unlawfully after countless hours of police and prosecutor time. Greater compliance with law also improves police-community relations.

And as with any unlawful search or seizure by police, community relations are impaired—especially when they involve racially charged drug interdiction efforts.[8] And, if drugs are discovered, the time spent processing, investigating, and prosecuting the case will be lost, as any drugs discovered in a search subsequent to an illegal detention will be suppressed.[9] More accurate assessments of reasonable suspicion would

5. Rachel A. Harmon, *Federal Programs and the Real Costs of Policing*, 90 N.Y.U. L. REV. 870, 931 (2015) ("Traffic enforcement is both a traditional law enforcement activity and a useful means of discovering drug crimes."). Officers are permitted to search a car on the basis of probable cause and a dog sniff is only one way to establish probable cause. Illinois v. Caballes, 543 U.S. 405, 409 (2005) (observing approvingly that the trial court found that hit by a trained drug dog constitutes probable cause to search a vehicle for drugs). Probable cause then permits a search of the vehicle and its containers. California v. Acevedo, 500 U.S. 565, 580 (1991). Using a dog to sniff for drugs, however, seems to be a common occurrence in a drug interdiction stop. Beth A. Colgan, *Revenue. Race, and the Potential Unintended Consequences of Traffic Enforcement Reform*, 101 N.C. L. REV. 889, 917 (2023) (citing EVALUATION & INSPECTIONS, DIV. 17-02, U.S. DEP'T OF JUST., REVIEW OF THE DEPARTMENT'S OVERSIGHTS OF CASH SEIZURE AND FORFEITURE ACTIVITIES 20-21 (2017), https://oig.justice.gov/reports/2017/e1702.pdf) (observing that in a study of eighty-five drug interdiction seizures, drug dogs were used more than 90% of the time).

6. *See* Jordan Blair Woods, *Policing, Danger Narratives, and Routine Traffic Stops*, 117 MICH. L. REV. 635, 637 (2019) (observing that more officers die annually during traffic stops than motorists, but very seldom at the hands of the occupants of the vehicles they have stopped); Aaron R. Megar, *Road to Reform: The Case for Removing Police from Traffic Regulation*, 75 VAND. L. REV. EN BANC 13, 29-30 (2022) (describing physical violence and verbal abuse by police during traffic stops).

7. Jordan Blair Woods, *Traffic Without the Police*, 73 STAN. L. REV. 1471, 1519 (2021) ("[A] common precursor to traffic stops escalating into violence against officers was the invocation of police authority in some way during the stop beyond asking for basic information, requesting documentation, or running a records check.").

8. *See generally* Samuel R. Gross & Katherine Y. Barnes, *Road Work: Racial Profiling and Drug Interdiction on the Highway*, 101 MICH. L. REV. 651, 687-95 (2002) (describing frequently mentioned concerns about racial discrimination in drug interdiction).

9. 4 WAYNE R. LAFAVE, SEARCH AND SEIZURE: A TREATISE ON THE FOURTH AMENDMENT § 9.3(f) (2022) (stating that if a dog sniff occurs after "the time had run out on the traffic stop detention either because its immediate lawful objectives had been accomplished or they had not been accomplished because of stalling . . . then the dog sniff and its fruits are all suppressible consequences of the illegal detention, unless of course the sniff/search was consented to or the continuation of the detention beyond

therefore decrease unnecessary safety risks to officers and motorists, improve police-community relations, and better allocate officers' time and resources.

The reasonable suspicion standard is, however, difficult to evaluate because, as it is a totality of the circumstances test, anything an officer finds to be suspicious during a stop can be considered.[10] Courts have given officers this vague standard despite the fact that the Supreme Court has frequently noted that rules regulating police officers need to be clear so that officers, who often have to act quickly in dangerous circumstances, know the limits of their authority.[11] Like other multi-factored tests, reasonable suspicion provides little in the way of guidance to those whose activities are governed by the reasonable suspicion standard.[12] Data and modern computing power, however, may better inform police, prosecutors, defense counsel, and judges about such standards—at least when a large number of decisions have considered the application of multi-factored tests to a variety of factual scenarios.[13]

---

its otherwise lawful limits was justified by the existence of reasonable suspicion"); L. Timothy Perrin et al., *It Is Broken: Breaking the Inertia of the Exclusionary Rule*, 26 PEPP. L. REV. 971, 983-85 (1999) (describing lost time of officers and others when evidence is suppressed).

10. Michael L. Rich, *Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment*, 164 U. PA. L. REV. 871, 895-901 (2016) (contending that because reasonable suspicion, at least in theory, requires a consideration of the totality of the circumstances, something that the author contends is beyond the ability of an algorithm). As we illustrate below, however, in the context of drug interdiction stops, a finite and relatively small number of factors may explain the outcome in most all drug interdiction cases.

11. *See* Aidan Taft Grano, *Casual or Coercive? Retention of Identification in Police-Citizen Encounters*, 113 COLUM. L. REV. 1283, 1300-02 (2013) (observing that courts frequently endorse bright-line rules for traffic stops); Daniel T. Gillespie, *Bright-Line Rules: Development of the Law of Search and Seizure During Traffic Stops*, 31 LOY. U. CHI. L. REV. 1, 3 (1999) (observing that with bright-line rules "[p]olice officials can more easily instruct officers in broad, clear-cut terms as to the legal procedures for conducting searches and seizures"). Reasonable suspicion is perhaps a particularly good example of a vague test because courts have yet to even offer a definition of the legal standard. Alabama v. White, 496 U.S. 325, 330 (1990) ("Reasonable suspicion is a less demanding standard than probable cause not only in the sense that reasonable suspicion can be established with information that is different in quantity or content than that required to establish probable cause, but also in the sense that reasonable suspicion can arise from information that is less reliable than that required to show probable cause.").

12. In this way, reasonable suspicion would not qualify as law under Justice Holmes' perspective. Oliver Wendall Holmes, Jr., *The Path of the Law*, 10 HARV. L. REV. 457, 460-61 (1897) (describing predictability as the fundamental characteristic of law). *See also* Patrick M. McFadden, *The Balancing Test*, 29 B.C. L. REV. 585, 643-649 (1988) (observing inconsistency and lack of predictability in balancing tests). Courts have also often observed the lack of predictability inherent in multi-factor balancing tests. *See* Michigan v. Bryant, 562 U.S. 344, 393 (2011) (Scalia, J., dissenting).

13. Mathematical methods have been employed in other contexts to predict legal outcomes from the facts of cases. *See, e.g.*, Kevin D. Ashley & Stephanie Brüninghaus, *Automatically Classifying Case Texts and Predicting Outcomes*, 17 A.I. & L. 125, 125-28 (2009) (trade secrets); Katie Atkinson & Trevor Bench-Capon, *ANGELIC II: An Improved Methodology for Representing Legal Domain Knowledge* 12-21 (Proc. 19th Int'l Conf. A.I. & L., 2023) (trade secrets); Barton Beebe, *An Empirical Study of the Multifactor Tests for Trademark Infringement*, 94 CALIF. L. REV. 1581 (2006) (trademark); Barton Beebe, *An Empirical Study of U.S. Copyright Fair Use Opinions, 1978-2005*, 156 U. PA. L. REV. 549 (2008) (copyright fair use); Hsuan-Lei Shao et al., *Factors Determining Child Custody in Taiwan After*

As described below, we have applied computational methods in a novel way to determine whether an officer has reasonable suspicion to continue to hold a motorist for further investigation in a drug interdiction stop. This first involves training a large language model to identify the sentences courts use to describe suspicious factors in drug interdiction stops. A variety of machine learning models are then employed to predict whether courts would find a particular combination of factors sufficient for reasonable suspicion.

In describing the current status of our work below, this Article starts with the legal problem: considering the feasibility and virtues of computational assessments of reasonable suspicion in drug interdiction stops. Despite the definition of reasonable suspicion as a totality of the circumstances test, practically, a very manageable number of commonly occurring factors determine whether officers may lawfully detain a car until a drug dog can be summoned.[14] Officers have, at most, a few minutes during a traffic stop to observe a car and its contents and interact with the driver and any passengers.[15] There are only so many different facts that an officer can identify in that time. Additionally, officers at the federal, state, and local levels are all trained to look for the same factors under a federal program that began during the War on Drugs in the 1980s and therefore they look for similar indicia of drug trafficking.[16] A manageable number of factors can therefore be modeled to assess reasonable suspicion.

A large data set exists of cases applying these factors. Legal challenges to drug interdiction stops occur frequently in state and federal courts.[17] It is estimated that 40,000 judicial opinions evaluating reasonable suspicion

Patriarchy's Decline: Decision Tree Analysis on Family Court Decisions, 17 ASIAN J. COMPAR. L. 272 (2022) (child custody); Allison Chorley & Trevor Bench-Capon, *AGATHA: Using Heuristic Search to Automate the Construction of Case Law Theories*, 13 A.I. & L. 9 (2006) (property rights in prey); Rafe Athar Shaikh et al., *Predicting Outcomes of Legal Cases Based on Legal Factors Using Classifiers*, 167 PROCEDIA COMP. SCI. 2393, (2020) (criminal law); L. Karl Branting et al., *Scalable and Explainable Legal Prediction*, 29 A.I. & LAW 213 (2021) (uniform domain name disputes).

14. This is, in some ways, a variation on the often-studied comparison between intuition and statistical analysis. Once a set of predictive factors is identified, a statistical analysis of those factors will more frequently predict the correct outcome than an individual's assessment of every circumstance observed in an individual case. *See* DANIEL KAHNEMAN, THINKING, FAST AND SLOW 222-33 (2011).

15. *See* Tracey Maclin, *Anthony Amsterdam's Perspective on the Fourth Amendment, and What It Teaches About the Good and Bad in* Rodriguez v. United States, 100 MINN. L. REV. 1939, 1968-83 (2016) (criticizing the Supreme Court for allowing questioning during traffic stop unrelated to officer's suspicion so long as traffic stop is not thereby prolonged).

16. David Rudovsky, *Law Enforcement by Stereotypes and Serendipity: Racial Profiling and Stops and Searches Without Cause*, 3 U. PA. J. CONST. L. 296, 300-01 (describing federal training of officers under Operation Pipeline and common practices of officers that followed that training).

17. *See, e.g.*, Rachel A. Harmon, *Federal Programs and the Real Costs of Policing*, 90 N.Y.U. L. REV. 870, 927-36 (2015) (describing dramatic increases in traffic enforcement as part of drug interdiction efforts as a result of federal funding for law enforcement as part of the War on Drugs).

are publicly available in electronic form.[18] Each of these cases is useful in developing a model of reasonable suspicion as the standard derives from the Fourth Amendment to the United States Constitution.[19] With courts at every level of judicial hierarchy, in every state in the nation, applying the same standard to a frequently occurring scenario in which a relatively small number of factors are considered, a computational model for reasonable suspicion is plausible. The weight courts in the aggregate assign to each factor should aid in predicting how any given court would view the variety of suspicious factors an officer observes.[20]

Such a model should improve the accuracy of reasonable suspicion determinations. Officers do not have access to these 40,000 cases while making stops along the side of the highway—or an analysis, of course, of these cases tailored to the facts the officer encounters.

Our early efforts suggest that the factors courts relied on in these 40,000 cases can be automatically identified, and that a predictive model that assesses reasonable suspicion can be developed. Somewhat understandably, there are fears about using machines to make legal decisions. The onus is certainly on those advocating computational approaches to, at a minimum, explain how such a proposal would work in a way that can be both understood and not oversimplified. Artificial intelligence and AI, though common phrases in any newscast, is still a mystery to many—and lawyers are no exception.[21] This Article intends to explain the processes used by the authors to identify the relevant information in the case law, and the application of the predictive models to the case law data, in a way that is accessible to lawyers with a genuine curiosity about machine learning but without a formal STEM background.[22]

To train a model to automatically identify the factors considered by courts in these 40,000 cases, the authors started with a sample of 211 drug interdiction cases from the Harvard Caselaw Access Project, annotated to identify the sentences that assess reasonable suspicion and the factors the

18. *See* Morgan A. Gray et al., *Toward Automatically Identifying Legally Relevant Factors*, *in* LEGAL KNOWLEDGE AND INFORMATION SYSTEMS 53, 54 (2022) (approximating number of cases).

19. Thomas Y. Davies, *The Supreme Court Giveth and the Supreme Court Taketh Away: The Century of Fourth Amendment "Search-and-Seizure" Doctrine*, 100 J. CRIM. L. & CRIMINOLOGY 933, 981-90 (2010) (describing development of reasonable suspicion as part of federalization of criminal procedure).

20. *Cf.* Mark A. Hall & Ronald F. Wright, *Systematic Content Analysis of Judicial Opinions*, 96 CALIF. L. REV. 63 (2008) (considering principal component analysis—without the assistance of computers—to assess case law involving multi-factor tests).

21. Curtis E. A. Karnow, *The Opinion of Machines*, 19 COLUM. SCI. & TECH. L. REV. 136 (2017) (introducing lawyers to the concept of machine learning).

22. Judge Karnow's excellent article was one of the earliest attempts to introduce lawyers to the concept of machine learning uses generic examples. *Id.* It is our hope that this Article will similarly introduce lawyers to these concepts while explaining the specifics our work on this particular issue.

courts relied on in reaching their conclusions.[23] We constructed a software "pipeline" to process these sentences. Based on the manually identified sentences and using large language models, it automatically identified similar sentences in other cases. With varying degrees of accuracy, these language models were able to identify whether a sentence was relevant to the reasonable suspicion calculation and the factor category in which it fit.

Our pipeline then has the ability to take factors and accurately predict reasonable suspicion. Various models were trained using the 211 cases with manually-assigned factors to see if the models could predict the outcome in test cases. Even the most basic machine learning models employed were able to correctly assess reasonable suspicion with greater than 80% accuracy, with some of the more complex models correctly assessing reasonable suspicion with 97.5% accuracy.

This Article attempts to explain the models, thus shedding light on the process of prediction by machine, at least in the case of our example. This introduction first attempts to allow the skeptical reader to understand how the models that we applied to our data work conceptually and gain some understanding of the math behind them.[24] Fear of statistical models, however, stems from more than just a lack of mathematical understanding. Even if one understands what a model is doing, it may not be clear how the model is using the data it uses to arrive at an output or prediction.[25] As models become more complicated, it can become difficult to discern the role, in this case, that the factors play in the assessment of reasonable suspicion. In fact, fearful descriptions of AI often refer to machine learning models as a "black box."[26]

To demystify machine learning, this Article explains three types of intuitive models: Logistic Regression, Decision Trees, and *k*-Nearest Neighbors. We used these models to make predictions, and the more

23. The work was conducted through the Center for Text Analytic Methods in Legal Studies, a research collaboration of experts from the University of Pittsburgh's Schools of Law and Computing and Information, the RAND Corporation, Duquesne Law School, and Worcester Polytechnic Institute, and supported by a Pitt Momentum Funds 2022 Scaling Grant. *See Center for Text Analytic Methods in Legal Studies*, UNIV. OF PITT. SCH. OF L., https://www.law.pitt.edu/center-text-analytic-methods-legal-studies (last visited July 24, 2023). Law students from the University of Pittsburgh School of Law received guidelines for annotating factors in automotive stop cases, a glossary of the relevant factors, and guided practice in annotating example cases.

24. This phenomenon has been referred to "algorithmic aversion." *See* Mirko Bagaric et al., *The Solution to the Pervasive Bias and Discrimination in the Criminal Justice System: Transparent and Fair Artificial Intelligence*, 59 AM. CRIM. L. REV. 95, 98 (2022).

25. Brandon L. Garrett & Cynthia Rudin, *The Right to a Glass Box: Rethinking the Use of Artificial Intelligence in Criminal Justice*, 109 CORNELL L. REV. 561, 586-604 (2024). (calling for AI devices that are explainable).

26. *See, e.g.*, FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015).

sophisticated devices were built on these. The fact that the principles behind these models are relatively easy to understand does not mean it is clear what the models are doing with the suspicious factors courts have considered to determine reasonable suspicion to predict the outcome in other cases. Of these three intuitive models, Linear Regression and Decision Trees have an advantage over *k*-Nearest Neighbors, as in these one can see how the model is handling each factor.[27] As more complexity is added to each of these models, however, their methodology becomes less easy to explain and the role factors play in the models' outputs becomes less clear.

Our experiments suggest that reasonable suspicion can be accurately predicted interpreted using an intuitive model, but only if the factors courts rely upon in a large corpus of judicial opinions can be reliably identified. It is often assumed that more complex models are better predictors of outcomes.[28] This is not always the case and when models are more easily understood, they are obviously preferred. We found an excellent success rate with an interpretable model. Two machine learning models were able to correctly identify the presence or absence of reasonable suspicion 97.5% of the time on a corpus of 211 cases annotated by humans. One of these results was obtained by a model that is frequently described as quite uninterpretable—a Neural Network. Neural Networks are sometimes said to be modeled after the complex working of the human brain and involve a highly complex collection of interconnected mathematical equations.[29] A Modified Decision Tree, which by contrast is very intuitively easy to understand and produces results that can be explained, achieved the same degree of accuracy.[30]

Understandably, concerns about using AI in the criminal justice system are not limited to what many find to be the mysteries of mathematical processes themselves—or the ability to understand how a model uses a

27. Professor Elyounes has observed that there is a natural preference for explainable models, especially in the legal context, and offers linear regression models as the best suited for this goal. Doaa Abu Elyounes, *Bail or Jail? Judicial Versus Algorithmic Decision-Making in the Pretrial System*, 21 COLUM. SCI. & TECH. L. REV. 376, 432 (2020).

28. *See, e.g.*, James Ming Chen, *Models for Predicting Business Bankruptcies and Their Application to Banking and Financial Regulation*, 123 PA. ST. L. REV. 735, 737-43 (2019) (observing that machine learning models that involve more complex combinations of basic models produce more accurate results in predicting bankruptcies).

29. *See, e.g.*, Rylan Schaeffer et al., *No Free Lunch from Deep Learning in Neuroscience: A Case Study Through Models of the Entorhinal Hippocampal Circuit* (Proc. of the 36th Conf. on Neural Info. Processing Sys., Nov. 2022).

30. A Neural Network and a Random Forest produced the best results. Random Forests are not as easily explained as the basic Decision Tree model from which it is derived, it is however far more interpretable than Neural Networks. *See, e.g.*, Dragutin Petkovic et al., *Improving the Explainability of Random Forest Classifier – User Centered Approach*, PAC. SYMP. BIOCOMPUTING 204 (2018).

factor.[31] In fact, the results of the models themselves have proved troubling. AI tools used in the criminal justice system have been shown to replicate human bias.[32] Thus, our data set presents a thorny problem of identifying and dealing with bias. This project, after all, uses judicial opinions to predict judicial behavior. Judges may well go about rendering their decisions in a biased manner, but their decisions are still the law, and courts are expected to follow precedent.[33] This Article discusses how our methods can be used as an aid to understanding bias in this domain. For example, these models, with careful interpretation, may reveal that judges rely on characteristics that disproportionately appear in certain segments of the population seized in drug interdiction stops, such as prior drug convictions or nervousness. This Article therefore proposes alerting the end-users—police departments, judges, advocates, and the public—to the fact that factors were relied upon by courts that may have skewed particular decisions, and perhaps even reveal how courts would rule had those factors not been credited in previous cases.

I. REASONABLE SUSPICION CAN BE MODELED

Police officers, judges, and litigants would particularly benefit from an automated system that could predict judicial decisions under a vague multi-factor legal test. Others have discussed the possibility of developing and deploying automated suspicion analysis systems that rely on data to establish a factual basis for reasonable suspicion.[34] Our work seeks to employ machines to assess the legal status of the facts an officer identifies, rather than assemble facts from which a case for reasonable suspicion could be made.

The Supreme Court has frequently recognized the virtue of clear rules governing police officers' actions,[35] yet the vagueness of probable cause

31. *See* Theresa A. Gabaldon, *Doing the Numbers: The Numerate Lawyer and Transactional Law*, 3 AM. U. BUS. L. REV. 63 (2014) (discussing lawyers' aversion to math and proposals to overcome it).

32. *See, e.g.*, Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/4G83-MDAS] (describing racially discriminatory effects of a bail algorithm).

33. *See, e.g.*, Anthony T. Kronman, *Precedent and Tradition*, 99 YALE L. J. 1029 (1990).

34. *See, e.g.*, David Rudovsky & David A. Harris, *Terry Stops and Frisks: The Troubling Use of Common Sense in a World of Empirical Data*, 79 OHIO ST. L.J. 501 (2018); David Lehr & Paul Ohm, *Playing with Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 658-62 (2017); Andrew Guthrie Ferguson, *Big Data and Predictive Reasonable Suspicion*, 163 U. PA. L. REV. 327 (2015); Rich, *supra* note 10.

35. *See, e.g.*, Riley v. California, 573 U.S. 373, 398, 401 (2014) (observing the Court's "general preference to provide clear guidance to law enforcement through categorical rules"); Michigan v. Summers, 452 U.S. 692, 704-05, 705 n.19 (1981) (extoling virtues of "workable rules . . . on a categorical – not in an ad hoc, case-by-case fashion"). *See also* Nicholas A. Kahn-Fogel, *Probabilistic Presumptions*

and reasonable suspicion—arguably the standards police use most frequently[36]—is frequently noted by the Court and commentators.[37] With probable cause, an officer may arrest a suspect, search a car, or seize an unknown substance for testing in a forensic laboratory.[38] Reasonable suspicion requires a lesser threshold, and only permits the officer to temporarily hold a person, or property, so that a further investigation can confirm or dispel a suspicion of a crime.[39]

These standards are vague because each requires a "totality of the circumstances" analysis.[40] An officer may consider anything that makes it more likely that criminal activity is afoot. Just as the Supreme Court has recognized the need for clear rules, it has also recognized the necessity of a flexible rule for assessing suspicion.[41] With an open universe of possible

in Fourth Amendment Decision-Making, 59 HOUS. L. REV. 313, 314 (2021) (observing that "in its Fourth Amendment jurisprudence as a whole, the Supreme Court has vacillated between a commitment to clear rules to provide officers with definitive guidance on the legality of their conduct and rejection of such rules in favor of case-by-case, totality of the circumstances analysis"); Albert W. Alschuler, *Bright Line Fever and the Fourth Amendment*, 45 U. PITT. L. REV. 227, 229 (1984) (observing and criticizing Supreme Court focus on "concern about the lack of legal guidance afforded police officers"); Wayne R. LaFave, *"Case-by-Case Adjudication" Versus "Standardized Procedures": The Robinson Dilemma*, 1974 SUP. CT. REV. 127, 141-42 (arguing that search and seizure doctrines are "primarily intended to regulate the police in their day-to-day activities and thus ought to be expressed in terms that are readily applicable by the police in the context of the law enforcement activities in which they are necessarily engaged").

36. *See* Andrew Manuel Crespo, *Probable Cause Pluralism*, 129 YALE L.J. 1276, 1279 (2021) ("[T]he requirement to demonstrate probable cause – or its junior partner, reasonable suspicion – constitutes the core substantive constraint on police power in the United States." (citing Alschuler, *supra* note 35, at 243; then citing Dunaway v. New York, 442 U.S. 200, 213 (1979))).

37. *See* Ric Simmons, *Quantifying Criminal Procedure: How to Unlock the Potential of Big Data in Our Criminal Justice System*, 2016 MICH. ST. L. REV. 947, 950 (describing reasonable suspicion and probable cause as "imprecise and subjective" legal standards that could be improved using big data algorithms which "can be structured so that they are truly race neutral and take into account individualized conduct"); *see also*, KEVIN D. ASHLEY, ARTIFICIAL INTELLIGENCE AND LEGAL ANALYTICS 73 (2017) ("Since legal rules employ terms that can be vague and open-textured, a computational model of reasoning with cases would help.").

38. United States v. Watson, 431 U.S. 411 (1976) (holding probable cause alone permits an arrest in public); Carroll v. United States, 267 U.S. 132 (1925) (holding probable cause to believe contraband in car is sufficient to search the car), Texas v. Brown, 460 U.S. 730 (1983) (holding probable cause required to believe item being seized is evidence of a crime).

39. *See* Anthony G. Amsterdam, *Perspectives on the Fourth Amendment*, 58 MINN. L. REV. 349, 414 (1974) (referring to reasonable suspicion as the "pint-sized version of probable cause").

40. 4 BARBARA E. BERGMAN ET AL., WHARTON'S CRIMINAL PROCEDURE § 25:6 (14th ed. 2023); WAYNE R. LAFAVE, SEARCH AND SEIZURE: A TREATISE ON THE FOURTH AMENDMENT § 9.5(b) (6th ed. 2024) (defining reasonable suspicion).

41. Frequently the critics of the vagueness of probable cause argue that some effort to express the certainty required for the standard should be expressed in mathematical terms. *See* Erica R. Goldberg, *Getting Beyond Intuition in the Probable Cause Inquiry*, 17 LEWIS & CLARK L. REV. 789 (2013); Ronald J. Bacigal, *Making the Right Gamble: The Odds on Probable Cause*, 74 MISS. L.J. 279, 338 (2004) (advocating a "tiered model of the levels of certainty required for searches and seizures"). *But see* Orin Kerr, *Why Courts Should Not Quantify Probable Cause*, *in* THE POLITICAL HEART OF CRIMINAL PROCEDURE: ESSAYS ON THEMES OF WILLIAM J. STUNTZ 131, 131-32 (Michael Klarman et al. eds., 2012). The role of the severity of the crime has occasionally been proposed as part of the probable cause calculus but only very rarely have judges entertained the possibility that this ought to be an express factor.

considerations, no single court decision could provide guidance to officers, judges, or litigants trying to determine whether there exists probable cause or reasonable suspicion.[42] In any given case, judges identify whether a particular set of facts meets one of the standards but usually they cannot explain how cases with different facts would be resolved.[43] As the Court has recognized, "there are so many variables in the probable-cause equation that one determination will rarely be useful 'precedent' for another."[44] Officers are, in many ways, left on their own to decide whether the circumstances they encounter are sufficiently suspicious to satisfy these legal standards.[45] If the law is going to have threshold standards like probable cause and reasonable suspicion, this lack of clarity is perhaps inevitable—at least with the limits on human decision-makers.

Modern computers may, however, have the capacity to improve human actors' ability to predict how a court would resolve a particular case under a multi-factor test, or a totality of the circumstances test like probable cause or reasonable suspicion. It is at least conceivable that a machine could identify the facts that courts have considered in previous decisions and whether those facts met the legal standard at issue. If there are enough such cases, seemingly a model could be designed that identifies the extent to which a factor by itself, and in conjunction with other factors, increases the likelihood that the standard has been satisfied.

A. The Law of Drug Interdiction Stops

Our first effort to create a computational model of suspicion considers a manageable subset of police-citizen encounters in which officers made a judgment about whether a crime has been committed. In drug interdiction stops, officers must decide—hundreds of times each day—

See Anne Bowen Poulin, *The Fourth Amendment: Elusive Standards, Elusive Review*, 67 CHI.-KENT L. REV. 127 (1992) (describing an en banc Seventh Circuit case which asserted that probable cause varied depending on the severity of the crime); Alschuler, *supra* note 35, at 247-48 (observing that Justice Jackson's opinion in Brinegar v. United States, 338 U.S. 160, 183 (1949) (Jackson, J., dissenting) was the only proposal to calibrate probable cause based on the seriousness of the crime but this was published before Judge Posner's Seventh Circuit opinion).

42. Craig S. Lerner, *The Reasonableness of Probable Cause*, 81 TEX. L. REV. 951, 953 (2003) (describing probable cause as "that elusive and perhaps hopelessly indeterminate constitutional standard").

43. Kit Kinports, *The Dog Days of Fourth Amendment Jurisprudence*, 108 NW. U. L. REV. COLLOQUY 64, 64 (2013) ("Totality-of-the-circumstances inquiries can be messy and unpredictable . . . .").

44. Illinois v. Gates, 462 U.S. 213, 238 n.11 (1983). *But see* Kahn-Fogel, *supra* note 35. (observing that officers are entitled a sufficient number of presumptions in their assessment of probable cause to give the standard aspects of a bright-line rule).

45. Anna Lvovsky, *The Judicial Presumption of Police Expertise*, 130 HARV. L. REV. 1995, 2044-52 (2017) (comparing the vagueness of probable cause to criminal statutes declared void for vagueness).

whether they have reasonable suspicion to detain a car pulled over for an ordinary traffic offense until a trained dog can sniff for the presence of drugs.[46] While the officer may consider anything when evaluating the likelihood that narcotics are present in such a situation, often only a limited number of factors prompt officers to detain cars.[47] In each traffic stop, an officer encounters a similar scene—there are only so many different things that can appear in an automobile. Additionally, the Fourth Amendment imposes a myriad of limits on traffic stops, giving officers a short time frame to identify any clear signs of drug trafficking. Finally, the frequency of such stops provides a relatively large corpus of judicial decisions considering approximately twenty bases of suspicion.[48]

The War on Drugs in the mid-1980s included a common national officer training that may have played a role in the small number of factors identified as supporting a claim of reasonable suspicion. In 1986, the Drug Enforcement Agency ("DEA") began to train their agents, as well as state and local police departments, on highway drug interception.[49] The program, called Operation Pipeline, instructed officers to find a legal way to stop suspected motorists.[50]

There was little need to train officers on how to stop a car. The law then imposed—and now imposes—few barriers on this first step.[51] An officer has extraordinary power to seize any motorist traveling on a highway.[52] While the Fourth Amendment requires that an officer believes a motorist is guilty of at least a traffic infraction to stop the car,[53] judges

46. *See* Michael C. Gizzi, *Pretextual Stops, Vehicle Searches, and Crime Control: An Examination of Strategies Used on the Frontline of the War on Drugs*, 24 CRIM. JUST. STUD. 139 (2011); Robert Massbarger, *Analysis of Traffic Stops Involving Drug Seizures*, 21 UNDERGRADUATE RSCH. J. 1, 12 (2017).

47. *See* KAHNEMAN, *supra* note 14, at 222-33 (describing greater reliability of statistical than intuitive or clinical predictions).

48. FRANK R. BAUMGARTNER ET AL., SUSPECT CITIZEN: WHAT 20 MILLION TRAFFIC STOPS TELL US ABOUT POLICING AND RACE 78-94 (2018) (describing statistics in North Carolina study explaining the types of police interactions that follow the initiation of a routine traffic stop).

49. Jane Bambauer, *Hassle*, 113 MICH. L. REV. 461, 504-05 (2015) (citing David Kocieniewski, *New Jersey Argues that the U.S. Wrote the Book on Race Profiling*, N.Y. TIMES, Nov. 29, 2000, at A1).

50. *See* Ricardo J. Bascuas, *Fourth Amendment Lessons from the Highway: A Principled Approach to Suspicionless Searches*, 38 RUTGERS L. J. 719, 761-69 (2007) (describing Operation Pipeline).

51. *See* Tracey Maclin, *Race and the Fourth Amendment*, 51 VAND. L. REV. 333 (1998) (describing the vast powers officers have to stop motorists and investigate the possibility that the motorist is transporting drugs).

52. *See, e.g.*, Devon W. Carbado, *From Stopping Black People to Killing Black People: The Fourth Amendment Pathways to Police Violence*, 105 CALIF. L. REV. 125, 130, 131-46 (2017) (describing the "broad discretion police officers have to force race-based interactions with African Americans without triggering the Fourth Amendment").

53. *See* Delaware v. Prouse, 440 U.S. 648 (1979) (holding stopping a car absent reason to suspect wrongdoing is forbidden). Officers are also permitted to stop cars in the absence of a traffic violation as part of a roadblock, but only if there is a non-criminal primary purpose for the roadblock such as

have frequently observed that virtually all motorists on the road will violates at least one motor vehicle law each time they drive.[54]

Operation Pipeline did, however, provide a great deal of detail regarding reasonable suspicion once a motorist was stopped. The DEA instructed officers to look for factors indicative of drug trafficking like: (1) fearing interaction with police; (2) attempting to conceal true travel plans from officers; (3) traveling in unusual ways, such as taking short roundtrips or driving for extended periods of times; (4) trying to conceal the odor of drugs; (5) making efforts to make space in the vehicle where items could be concealed; (6) traveling on known drug corridors; or (7) driving a vehicle not owned by the driver or any passenger.[55] Not surprisingly given that officers were all trained to look for similar types of suspicious circumstances, the facts officers use to support their claims that reasonable suspicion justifies continued detention almost always fit into these criteria.

The limited time of a roadside encounter further imposes a practical limit on an officer's ability to investigate a wide range of possible bases of suspicion that fall outside these categories. As the Supreme Court has stated, the "detention of a motorist pursuant to a traffic stop is presumptively temporary and brief."[56]

While there is some variation in the laws regulating traffic stops across the country, there is also remarkable uniformity. For example, states have generally chosen not to draft their laws relating to traffic stops in a way that expands upon the Fourth Amendment's protections.[57] An officer may

interception drug drivers or illegal aliens. *See* United States v. Martinez-Fuerte, 428 U.S. 543 (1976) (permitting checkpoint for aliens fifty miles from the U.S./Mexico border); Michigan Dep't of State Police v. Sitz, 496 U.S. 444 (1990) (permitting DUI checkpoint); City of Indianapolis v. Edmond, 531 U.S. 32 (2000) (holding a roadblock to discover drug possession and trafficking unconstitutional).

54. The Court has long recognized that any traffic offense permits a stop. *See Prouse*, 440 U.S. 648. More recent decisions have concluded that even if few officers would ever stop such a motorist for the offense in question, that the mere commission of any traffic offense is sufficient. *See* Whren v. United States, 517 U.S. 806, 810 (1996) (petitioners argued that "the use of automobiles is so heavily and minutely regulated that total compliance with traffic and safety rules is nearly impossible, a police officer will almost invariably be able to catch any given motorist in a technical violation"). *See also* David A. Sklansky, *Traffic Stops, Minority Motorists, and the Future of the Fourth Amendment*, 1997 SUP. CT. REV. 271, 273 ("Since virtually everyone violates traffic laws at least occasionally . . . police officers, if they are patient, can eventually pull over anyone they choose."). On the interstate highway, it is difficult to find motorists who are obeying the speed limit and those who do so may not be driving safely. *See* Gregory H. Shill, *Should Law Subsidize Driving*, 95 N.Y.U. L. REV. 498, 506-11 (2020) (observing culture of disrespecting speed limits and lax enforcement of speeding); Benjamin I. Schimelman, *How to Train a Criminal: Making Fully Autonomous Vehicles Safe for Humans*, 49 CONN. L. REV. 327, 342-43 (2016) (observing that training autonomous vehicles to obey speed limits would often make them unsafe because they are traveling slower than the flow of traffic).

55. Bambauer, *supra* note 49, at 504-05.

56. Berkemer v. McCarty, 468 U.S. 420, 437 (1984). *See also Prouse*, 440 U.S. at 653 (holding "resulting detention" from a traffic stop should be "quite brief").

57. *See, e.g.*, Margaret M. Lawton, *The Road to* Whren *and Beyond: Does the "Would Have" Test*

observe the vehicle and speak to the driver and any passenger about any subject while performing their permissible, legal tasks during the stop.[58] Every United States jurisdiction permits officers to ask motorists for their driver's license and registration.[59] Some jurisdictions allow officers to take time to ask the driver, and even the passenger, about travel plans during the routine stop; but other jurisdictions find this beyond the permissible scope of an ordinary traffic offense.[60] Either by entering the information from these documents into an electronic system, or by reporting their contents to a dispatcher, an officer may attempt to determine whether the documents are valid and whether there are any outstanding warrants for the driver. The vehicle can be detained while the officer awaits the answer to this query, so long as the response time is not unreasonably long.[61] Typically, this information can be determined in a couple of minutes or less.[62] In some jurisdictions, the officer may also detain the car until the driver's and any passenger's criminal records are

Work?, 57 DEPAUL L. REV. 917, 956-58 (2008) (observing that even when state constitutional criminal procedure provides greater protections for motorists, practically speaking, because of the complexity of criminal procedure relating to traffic stops, officers have very similar powers).

58. Muehler v. Mena, 544 U.S. 93, 100-01 (2005) (holding questioning during an investigatory detention is not limited to matters that led to the detention so long as questioning does not extend the length of the stop); Arizona v. Johnson, 555 U.S. 323, 333 (2009) (same).

59. *See, e.g.*, United States v. Henley, 469 U.S. 221, 235 (1985) (stating that officers conducting a traffic stop are permitted to "take such steps as [are] reasonably necessary to protect their personal safety and to maintain the status quo during the course of the stop"); United States v. Shabazz, 993 F.2d 431, 437 (5th Cir. 1993) (during traffic stop, officer may request driver's license, insurance papers, vehicle registration, and run computer records check); West v. United States, 100 A.3d 1076, 1085-86 (D.C. Ct. App. 2014) (officer may look inside vehicle during traffic stop).

60. There is a conflict in the courts over whether an officer may routinely ask the driver about travel plans without reasonable suspicion of criminal activity beyond the initial traffic offense. In every jurisdiction it is permissible, so long as the questions do not extend the length of the stop. In most jurisdictions that considered this issue, questions about travel plans are a part of the permissible routine questions and do not extend the time of the stop regardless of when they are asked. *See* Carlisle v. Commonwealth, 601 S.W.3d 168, 177 (Ky. 2020); United States v. Campbell, 912 F.3d 1340, 1354 (11th Cir. 2019) ("Generally, questions about travel plans are ordinary inquiries incident to a traffic stop."); United States v. Cole, 21 F.4th 421 (7th Cir. 2021) (en banc) (reversing divided panel decision that concluded questions about travel plans were impermissible and thus extended the length of the stop, with three judges dissenting from this conclusion in the en banc decision); United States v. Williams, 271 F.3d 1262, 1267(10th Cir. 2001); United States v. Brigham, 382 F.3d 500, 507-08 (5th Cir. 2004). A minority of courts to consider the issue, however, have resolved that such questions may only be asked if they do not lengthen the stop, for instance while the officer is awaiting the report on the validity of the driver's license. State v. Jimenez, 420 P.3d 464, 476 (Kan. 2018) (police officer's questions to defendant and passenger during traffic stop regarding their travel plan impermissibly prolonged stop).

61. Court have frequently stated that the length of time for this records check must not be unreasonably long but have been reluctant to provide any meaningful parameters on the upper boundary of the length of time it is appropriate to wait. *See, e.g.*, Byndloss v. State, 893 A.2d 1119, 1132-34 (2006) (computer problems that delay records check by twenty minutes does not unreasonably extend the stop).

62. In 1979, the California Attorney General claimed in a brief that this process should take at longest four minutes but can sometimes be completed in a few seconds. *See* Wayne R. LaFave, *The "Routine" Traffic Stop from Start to Finish: Too Much "Routine," Not Enough Law*, 102 MICH. L. REV. 1843, 1877 (2004) (citing People v. McGaughran, 601 P.2d 207, 211 n.6 (Cal. 1979)).

determined.[63]

While conducting these activities, officers are permitted to visually inspect anything in the car and question the driver and any passenger about any topic, so long as these actions do not extend the length of the stop.[64] There is thus a very limited window to gather information that might suggest the motorist is transporting drugs. To gather information, an officer may examine the contents of the car and the occupants' demeanor while they process the driver's documents. Additionally, an officer may extend a stop slightly by asking the driver and any passenger to exit the vehicle in the interest of officer safety, even if nothing suggests that there is a threat posed by those stopped.[65] In jurisdictions that regard questions about travel plans to be part of an officer's routine investigation in any traffic stop, the officer may ask questions about travel plans even if those questions extend the length of the traffic stop. In other jurisdictions, such questions are still permissible so long as they do not extend the stop, for instance if asked while the officer is waiting for a dispatcher to respond to a request to determine the status of the driver's license. This period, from the stop to the response to the records check, however, is very brief, typically lasting less than five minutes.[66]

63. Courts are divided on whether an officer is allowed to determine the criminal history of the driver and any passengers without extending the length of the stop. *See Carlisle*, 601 S.W.3d at 177 (identifying cases that are dividing federal appellate courts); United States v. Hill, 852 F.3d 377, 383 (4th Cir. 2017) ("[A]n officer reasonably may search a computer database during a traffic stop to determine an individual's prior contact with law enforcement, just as an officer may engage in the indisputably proper action of searching computer databases for an individual's outstanding warrants."); State v. Allen, 779 S.E.2d 248, 257-58 (Ga. 2015) (permitting criminal history check for driving and passenger as part of routine traffic stop in the interest of officer safety); *but see* United States v. Evans, 786 F.3d 799, 786 (9th Cir. 2015) (holding that running an "ex-felon registration check" was "wholly unrelated" to a traffic stop for an ordinary traffic violation).

64. *West*, 100 A.3d at 1085-86; Muehler v. Mena, 544 U.S. 93, 100-01 (2005); Arizona v. Johnson, 555 U.S. 323, 333 (2009).

65. Pennsylvania v. Mimms, 434 U.S. 106 (1977) (driver may be ordered out of car as part of routine traffic stop in the interest of officer safety); Maryland v. Wilson, 519 U.S. 408 (1997) (passengers may also be ordered out of car during stop in the interest of officer safety).

66. *See* LaFave, *supra* note 62, at 1877. Interestingly, officers do not appear to have taken advantage of a type of loophole in the regulation of traffic stops that existed for nine years that would have given officers extraordinary discretion to end the length of traffic stops with no suspicion beyond the rationale for the initial stop. With the Supreme Court's decision in *Atwater v. City of Lago Vista*, 532 U.S. 318, 354 (2001), the Court held that a police officer may lawfully arrest a motorist for even the most minor of traffic offenses. *Virginia v. Moore*, 553 U.S. 164, 176 (2008) went even further, holding that even if a state legislature forbade a custodial arrest for a minor offense, evidence discovered incident to that arrest would not be subject to exclusion under the federal exclusionary rule. Until the Court's decision in *Arizona v. Gant*, 556 U.S. 332 (2009), the Court's search-incident-to-arrest doctrines allowed an officer to search a car merely by choosing to arrest. Even more troublesome, the Court's decision in *Rawlings v. Kentucky*, 448 U.S. 98, 111 (1980) allowed an officer to perform the search-incident-to-arrest before the arrest, meaning that actually taking a motorist into custody was not a pre-requisite to the search. Probable cause to believe the motorist guilty of a minor traffic offense was sufficient to search the car if the officer merely claimed an arrest was planned, a plan that could be aborted if the search was fruitless. *Gant* changed the Court's search-incident-to-arrest doctrine permitting the search of a car only if the arrest was

Though courts permit officers to consider anything they discover and deem suspicious during these stops, in this limited time, it is unlikely that there will be an indicator of the presence of drugs that has not been considered by other officers and evaluated by many courts. There are, after all, only so many different things that can be in a vehicle and only so much information that can be learned from the short interview permitted. Most of the time, officers identify combinations of factors from this relatively short list of factors to support their belief that drugs are present in the car.

B. The Ill-Defined Reasonable Suspicion Standard

It is far from clear whether any particular set of factors lawfully allows an officer to detain a motorist for further investigation.[67] Once a car is stopped for an ordinary traffic offense, it may be detained for a sniff by a drug dog only if the indicators rise to the level of "reasonable suspicion."[68] Like probable cause, this standard is quite vague. By contrast, however, courts have at least arrived at a definition of probable cause—a quantum of evidence "sufficient . . . to warrant a [person] of reasonable caution in the belief that an offense has been or is being committed"[69]—however unhelpful that definition is in assessing whether a particular set of facts satisfies the standard. A search for even a statement of the definition of reasonable suspicion in case law proves elusive.

for an offense that might yield evidence. *See* Seth W. Stoughton, *Modern Police Practices:* Arizona v. Gant*'s Illusory Restriction of Vehicle Searches Incident to Arrest*, 97 VA. L. REV. 1727 (2011) (recognizing *Gant* had reduced potential for suspicionless searches but noting other doctrines that give officers broad authority to search). If a motorist is arrested for an offense of any sort, a motorist's car may still be searched without suspicion, but only as part of an inventory search if the car is impounded, which assumed that the motorist was actually taken into custody and booked. *See* Colorado v. Bertine, 479 U.S. 367 (1987); South Dakota v. Opperman, 428 U.S. 364 (1976).

67. *See* United States v. Garrido, 467 F.3d 971, 982 (6th Cir. 2006) (observing that Sixth Circuit precedents involving determination of whether an officer has reasonable suspicion to detain a motorist for a drug dog are "inconclusive" in assessing whether reasonable suspicion is present in cases involving different facts).

68. Rodriguez v. United States, 575 U.S. 348 (2015) (reasonable suspicion required to detain motorist after conclusion of the purposes for the stop).

69. Heien v. North Carolina, 574 U.S. 54, 77 (2014) (Sotomayor, J., dissenting); Safford Unified School District #1 v. Redding, 557 U.S. 364, 370 (2009); Draper v. United States, 358 U.S. 307, 313 (1959); Brinegar v. United States, 338 U.S. 160, 175-76 (1949); Carroll v. United States, 267 U.S. 132, 162 (1925). *See* Davies, *supra* note 19, at 967 (tracing the history of this formulation to a jury instruction given by U.S. Supreme Court Justice Bushrod Washington in *Munns v. De Nemours*, 17 F. Cas. 993, 995 (C.C. Pa. 1811) while riding circuit).

At no point has a court offered a clear, concise definition of reasonable suspicion.[70] Though the term "reasonable suspicion" was fashioned by the Supreme Court in 1968,[71] the Court in 1981 observed that "[c]ourts have used a variety of terms to capture the elusive concept."[72] The Court has recognized the difficulty of even stating the reasonable suspicion test, noting that it "turn[s] on the assessment of probabilities in particular legal factual contexts — not readily, or even usefully, reduced to a neat set of legal rules."[73] In one formulation by the Court, reasonable suspicion has been satisfied if "specific articulable facts, together with rational inferences from those facts . . . reasonably warrant suspicion" that a particular crime has occurred or is occurring.[74] The Court elaborated that reasonable suspicion is "some minimal level of objective justification"[75] that is "something more than an 'inchoate and unparticularized suspicion or hunch.'"[76] Without offering to ballpark the mathematical level of certainty required, the Court indicated that the bar is not high, noting that "the reasonable suspicion inquiry 'falls considerably short' of 51% accuracy."[77]

The Court has further (candidly) offered that the reasonable suspicion standard "fall[s] short of providing clear guidance dispositive of the myriad factual situations that arise."[78] In the context of drug interdiction stops, this means that case law interpreting reasonable suspicion alone provides an officer guidance on whether there is an adequate legal basis to detain a motorist until a drug dog can be summoned to the scene.[79] This is highly problematic for an officer, who must, in a short period of time, decide whether to release or hold the motorist for further investigation. The officer does not have access to a law library on the shoulder of the highway nor would there be an opportunity to comb through it if it were

70. Deborah Anthony, *The U.S. Border Patrol's Constitutional Erosion in the "100-Mile Zone"*, 124 PENN ST. L. REV. 391, 410 (2020); Sean A. Brown, *Terry Stops: Cracking the Code of Reasonable Suspicion*, 105 ILL. BAR J. 42 (2017) (stating there no clear definition of reasonable suspicion).

71. Terry v. Ohio, 392 U.S. 1 (1968); Jeffrey Fagan, Terry*'s Original Sin*, 2016 U. CHI. LEGAL F. 43 (2016) (observing *Terry* as origin of reasonable suspicion standard); *see also* Lvovsky, *supra* note 45, at 2044-52 (comparing the vagueness in the reasonable suspicion standard to loitering statutes that have been declared void for vagueness); Rudovsky & Harris, *supra* note 34, at 506; David A. Harris, *Particularized Suspicion, Categorical Judgments: Supreme Court Rhetoric Versus Lower Court Reality Under* Terry v. Ohio, 72 ST. JOHN'S L. REV. 975 (1998).

72. United States v. Cortez, 449 U.S. 411, 417 (1981). *See also* Amsterdam, *supra* note 39, at 410.

73. Illinois v. Gates, 462 U.S. 213, 232 (1983).

74. United States v. Brignoni-Ponce, 422 U.S. 873, 884 (1975).

75. INS v. Delgado, 466 U.S. 210, 217 (1984).

76. United States v. Sokolow, 490 U.S. 1, 7 (1989); *Terry*, 392 U.S. at 27.

77. Kansas v. Glover, 140 S. Ct. 1183, 1188 (2020).

78. United States v. Cortez, 449 U.S. 411, 417 (1981).

79. *See* Alschuler, *supra* note 35, at 287 (recommending teaching case-based reasoning in police academies to provide officers insight on how to deal with less than clear standards).

there. Although practically there are only so many different scenarios that an officer can encounter in a search for drug trafficking among ordinary traffic offenders, it would be a Herculean task—even for a court—to keep track of how courts in any jurisdiction have assessed each combination of suspicious circumstances.[80] Assuming an officer, judge, or any human for that matter, had this capacity, how can case law be helpful when the combination of suspicious factors the officer encounters has not yet been directly or clearly addressed by any court?

C. The Potential Benefits of an Automated Standard

Modern computing power is perhaps tailor-made to assist police officers, advocates, and judges interested in the likelihood that a combination of factors rises to the level of reasonable suspicion that drugs are present in a car.[81] If legal decisions exist involving the same factors an officer faced, traditional legal research tools would make it difficult for officers, or prosecutors who may be advising them, to summon those cases quickly. When cases with the same factors do not exist, machine learning can be used to shed light on what courts and officers should do in such a situation. This would require using all relevant cases to evaluate the weight courts give factors by themselves, and in combination with other factors, to predict how a court would view a particular set of factors. Certainly, lawyers look at related cases and, using common sense judgment, offer a sense of where courts are going with an issue. The number of useful cases that bear on reasonable suspicion in drug interdiction stops would certainly complicate this task for a hypothetical lawyer with limitless time—an impossible task for an officer with a minute on the shoulder of a highway to assess the legal landscape. Reading all these cases to develop fully formed intuition about how a court would resolve such an issue is beyond the capacity of a human lawyer—and, at a minimum, is not an effective use of a human lawyer's time.

Certainly though, it would be in the best interest of police departments, judges, litigants, and society more generally, to have a more legally accurate assessment of reasonable suspicion in drug interdiction stops.[82]

80. One legal commentator has observed that the ideal judge is just such a person, one with a Herculean ability to discover, retain, and process precedent. RONALD DWORKIN, LAW'S EMPIRE 239 (1986). In the context of reasonable suspicion, a machine can come much closer to this idealized judgement than a human.

81. *See* Simmons, *supra* note 3, at 950-51 (discussing use of machines to consider publicly available data and assess reasonable suspicion).

82. One consequence of an unlawful search is certainly the loss of evidence, which creates at least an optics concern for police who may be blamed for failing to obtain a conviction that would have

When officers illegally detain vehicles for a drug dog sniff, any drugs subsequently discovered are tainted by the unlawful seizure and are not admissible in a criminal proceeding.[83] In such situations, it may appear as though there is no loss to the police, as there would have been a lost conviction regardless of whether the officer searched the car unlawfully and found drugs that were inadmissible or allowed the car to continue without searching it. Considerable time, however, is lost when officers conduct unlawful searches. When drugs are discovered in a car, officers process the scene and book the suspects—time that is lost if an illegal detention leads to the suppression of the evidence. This time could have been lawfully spent investigating potential drug traffickers.[84] Furthermore, additional officer and prosecution time is spent on the case through the defense's successful hearing to suppress the drugs.

Far more substantial, however, is the effect of unsuccessful unlawful searches. Intrusions on the liberty, property, and privacy of persons that turn out to be unjustified violate defendants' rights and undermine police legitimacy in many communities.[85] In the context of drug interdiction, an unjustified search could be defined as one that is unlawful regardless of whether it yields drugs, or one that fails to yield drugs regardless of whether it is lawful. Both have a negative effect on police reputation, and either way, fewer illegal searches improve the situation, as it can be presumed that illegal searches are less likely to produce drugs than legal ones.[86] Current events demonstrate the need to improve trust in police.[87] Ineffective or illegal police work may damage police-community relations in communities where this relationship is the rockiest.

Since Operation Pipeline's inception, there have been claims that drug interdiction practices are racially discriminatory. Claims that Black motorists were more frequently stopped and searched led critics of the practice to identify a new "crime" called "Driving While Black."[88] While

otherwise occurred. *See, e.g.*, L. Timothy Perrin et al., *If It's Broken, Fix It: Beyond the Exclusionary Rule*, 83 IOWA L. REV. 669 (1998) (discussing historical frequency of success rates in motion to suppress).

83. *See, e.g.*, Rodriguez v. United States, 575 U.S. 348 (2015).

84. *See* Perrin et al., *supra* note 9, at 983-85.

85. *See, e.g.*, Tracey L. Meares et al., *Lawful or Fair? How Cops and Laypeople Perceive Good Policing*, 105 J. CRIM. L. & CRIMINOLOGY 297 (2015).

86. This, of course, assumes that searches performed with probable cause are more likely to be successful than those for which the officer lacks probable cause. This seems intuitively correct but the fact that legality and odds are not necessarily correlated has led some commentators to produce a search standard expressly based on probability. *See, e.g.*, Max J. Minzner, *Putting Probability Back into Probable Cause*, 87 TEX. L. REV. 913 (2009).

87. Lack of trust in some communities is certainly not a new phenomenon. *See, e.g.*, Richard R.W. Brooks, *Fear and Fairness in the City: Criminal Enforcement and Perceptions of Fairness in Minority Communities*, 73 S. CAL. L. REV. 1219 (2000).

88. Stephen Rushin & Griffin Edwards, *An Empirical Assessment of Pretextual Stops and Racial Profiling*, 73 STAN. L. REV. 637 (2021).

a Department of Justice report during George W. Bush's presidency revealed that minority motorists were no more likely than white motorists to be *stopped*, it did reveal that minority motorists were more likely to be *searched*.[89] It might be tempting to conclude that minority motorists do more frequently possess drugs and therefore more often appear suspicious. Statistical studies have demonstrated to the contrary, however, that searches of minority motorists are *less* likely to yield drugs, suggesting that police are simply more likely to search minority drivers' cars than non-minority drivers' cars.[90]

An automated system that could reliably determine whether a set of facts amounts to reasonable suspicion that a motorist is transporting drugs could reduce the number of illegal detentions, allow better allocation of police resources, reduce unjustified police intrusions, and improve community relations. This kind of assessment analysis may be feasible through techniques and methods in AI and machine learning. The number of possible considerations and combination of considerations are too numerous for humans to develop models but are manageable by computational methods. Police departments have all been trained to look for similar factors and, because they received the same training, tend to use similar language in describing these factors. Finally, courts throughout the nation considering these factors all apply the Fourth Amendment to the United States Constitution, so even the analyses of the various factors tend to use similar language.

To develop a predictive device from case law, however, requires access to a large corpus of judicial opinions interpreting the reasonable suspicion standard. Proprietary platforms, such as Westlaw and Lexis understandably charge for such access. Fortunately, the Harvard Caselaw Project provides open access to such a corpus and as discussed in the next Section, the factors and court holdings can—to a reasonable degree of accuracy—be identified in these decisions. With this data, a model can be created that predicts, based on the factors present, whether reasonable suspicion is present in each situation. Various models are possible, as discussed in Section III, with some showing real promise for consistently making an accurate assessment.

II. IDENTIFYING LEGALLY RELEVANT FACTORS WITH LANGUAGE MODELS

In Section III, we discuss using machine learning to predict and analyze case outcomes based on legally relevant factors. These factors, which we

89. Matthew J. Hickman, *Traffic Stop Data Collection Policies for State Police, 2004*, BUREAU OF JUST. STAT. FACT SHEET (June 2005), https://bjs.ojp.gov/content/pub/pdf/tsdcp04.pdf.

90. BAUMGARTNER ET AL., *supra* note 48, at 99-124.

explain in more detail below, describe the facts upon which a court reached its conclusion on the issue of reasonable suspicion to detain a motorist on suspicion of drug trafficking. Collecting these factors for use in machine learning models to predict and analyze case outcomes happens via hand annotation, which is time consuming and expensive. Ideally, we could use natural language processing to automatically identify what factors are present in an individual case. We discuss our efforts to automatically identify factors in case opinions below.

A. Developing a List of Suspicious Factors

A list of suspicious factors was created based on two methods. First, lawyers familiar with the legal issue read cases and identified that there were groups of factors relied on by courts in making the determination of whether suspicion was present. Second, we consulted the criteria the DEA developed in Operation Pipeline.[91] The table below identifies various specific factors officers were trained to identify as part of Operation Pipeline as signs of drug trafficking, some of which are obviously dated at the time of this writing.

91. *Cf.* Ashley & Bruninghaus, *supra* note 13, at 134 (describing starting with Restatement (First) of Torts to identify factors relevant in modeling of multi-factor trade secrets cases).

Table 1: Suspicious Factors Identified in DEA Operation Pipeline Training Materials[92]

Vehicle	**Motorists**
Luggage	Eyes
Too Much or Too Little for Trip	*The Window to the Soul*
Atlas/Map	Dilated Pupils
Route or Cities Marked Don't Match Story	*Especially During Relevant Questions*
Tool/Marks	Excessive Blinking
On Screws, Panel, Nuts, and Bolts	Eyes Wide Open
Key Rings	No Eye Contact
Note Only One Key – No Trunk Key	*Looks at Ground*
Fast Food Wrapper	Distant Look
Driver Doesn't Want to Leave Drugs or Money	Closes Eyes
CB Radio/Radar Detectors	*Hopes You Go Away*
CB Tuned to Low Use Frequency	Squints
Car Phone/Pager	Facial Twitches
Everyone Has Them	Dry Mouth
Odometer Reading	*Swallows Repeatedly, Licks Lips, Clears Throat*
High Mileage Suspicious on New Vehicle	Clenched Jaw
Odors	Frowning
Note Overwhelming Odors – Used to Mask Odor of Drugs	Twisted Mouth
Handles/Knobs	Yawning
Missing Door Handles/Knobs or Those Observed on the Floor May Indicate Contraband in Door	*Excessive or Repeated is a Strong Sign of Deception*
Newspapers	Bites Lips
City and Date Don't Match Story	Face Turns Red, Blushes
Tools	Face Pales, Turns White or Ashen
A Few Specific Tools May Indicate a Hidden Compartment	*Danger Signal*
Disclaimers	Sweating Inappropriate for Conditions
Police or Religious Stickers	Shaking Hands
Windows	General Body Tremors
If They Don't Roll Down They May Be Filled with Contraband	Hides Hands
Pre-Paid Phone Card	*Danger Signal*
Use of These Cards are Popular with Drug Courier	Points Away
Non-Factory Switches and Buttons	*Misdirection Signal*
May Activate Electric Compartment	Covers Eyes
Attorney Business Cards	Rubs or Touches Nose
May Indicate Preparedness to Contact Attorney in Case of Arrest	Plays with Mustache
Expensive Vehicle	Tugs at Ears
No Lien	Covers Ears
Tinted Window	Pats Cheek
Difficult to See In	*Reassuring Gesture*
Rental Vehicle	Pats or Smooths Hair
	Covers Mouth

92. This table is taken from training material created by a Kansas officer teaching methods of interacting with a motorist. Lt. Kirk Simone, *Epic Operation Pipeline: Passenger Vehicle Drug Interdiction*, https://norml.org/wp-content/uploads/pdf_files/brief_bank/Operation_Pipeline_Manual.pdf.

3rd Party Rental	*Doesn't Want You to See Him Lying*
Vehicle Registration	Covers Throat or Groin
3rd Party Owner Not Present	*Protective Gesture*
Duct Tape	Fidgets/Nervous Hands
Roll in Vehicle Could Have Been Used to Wrap or Seal Drugs	*Plays with Fingernails*
Trash Bags/Saran Wrap	*Toys with Jewelry*
Used to Package Drugs	Pats, Smooths, or Massages Any Part of Body
Paint/Bondo/Silicone/Adhesive	Taps Chest
Used to Cover or Seal Hidden Compartments or Vehicle Alterations	*Fingering the Culprit*
Screws	Rubs Hands or Fingers Together
Loose, Missing, Damaged, Non-Factory, Worn Heads	Scratches Repeatedly
Look Bent, Broken Interior Panels	Tugs at Clothing
May Indicate Use for Hiding Drugs	Continually Picks Lint
Signs of Recent Drug Use	Cannot Hold Arms or Hands Still
Roaches in Ashtray, Seeds, Needles, Rolling Papers, One Hitters, Bindles, Etc.	Folds Arms Across Chest
Motel Receipts	*Building a Barrier*
Do Dates and Locations Match the Driver's Story	Leans Excessively
	Locks Onto an Object
	Tense Rigid Movements
	Exaggerated Movements
	Restlessness
	Cannot Keep Feet Still
	Foot Tapping
	Nausea
	Vomiting
	Goose Bumps
	Hair Stands on End
	Arms and Back of Neck
	Pulse in Abdomen

We compared this list and with hundreds of Fourth Amendment decisions evaluating drug interdiction stops. To develop the final version of the list of factors, we looked to create a smaller list of more generic categories that encapsulated a number of the factors officers were instructed to look for. This final list of factors—that was used to annotate the cases—appears below in Table 2, with the accompanying number and letter identification scheme that was used in the annotation process.

Table 2: Factor Type System

1 Occupant Appearance or Behavior	**2 Occupant Status**	**3 Travel Plans**
1A Furtive Movement 1B Physical Appearance of Nervousness 1C Nervous Behavior 1D Suspicious or Inconsistent Answers	2E Motorist License 2F Driver Status 2G Refused Consent 2H Legal Indications of Drug Use 2I Motorist's Appearance Related to Drug Use	3J Possible Drug Route 3K Unusual Travel Plans
4 Vehicle	**5 Vehicle Status**	**6 Other Annotation Labels**
4L Expensive Vehicle 4M Vehicle License Plate or Registration 4N Unusual Vehicle Ownership	5O Indicia of Hard Travel 5P Masking Agent 5Q Vehicle Contents Suggest Drugs 5R Suspicious Communication Device 5S Suspicious Storage	6T Other 6U Possibly Off Point 6V Suspicion Found? – No 6W Suspicion Found? – Yes

This list was designed to create categories that meaningfully differentiate factors while minimizing overlap between the categories. It is worth nothing that at this point in our study, our pipeline of analysis has demonstrated a promising capacity to identify these categories in a judicial opinion; and using only the factors in Table 2, machine learning models, described more thoroughly in Section III, have been able to assess whether reasonable suspicion exists with, depending on the model, accuracy between 80 and 97.5%. More work remains to be done with the judicially used factor descriptions and the methods used to identify them, but these early results suggest the list of factors we use are a meaningful representation of court opinions.

Many factor names in Table 2 were taken directly from the language courts use in describing a factor as suspicious. Although each category, indicated in bold, may cover more than described herein, our descriptions cover the general area encompassed by each factor. Our first category, **Possible Drug Route** involves traveling on any highway that the court describes as one known for the transportation of drugs, such as Interstate 95 from Miami to New York.[93] The category **Unusual Travel Plans** can involve a variety of activities, but is typified by very short round trips, or the use of a rental car to travel a long distance.[94] A driver or passenger

93. United States v. Williams, 808 F.3d 238, 252 (4th Cir. 2015) (describing I-95 as a known drug corridor).

94. *See, e.g.*, United States v. Simpson, 609 F.3d 1140, 1151 (10th Cir. 2010) (describing long drive to Reno to spend one night in a friend's house to gamble to be sufficiently unusual to contribute to reasonable suspicion analysis).

offers **Suspicious or Inconsistent Answers** when the officer recounts the suspect's effort to avoid being candid—this can manifest itself in a statement by the driver that is itself nonsensical or is inconsistent with the passenger's account.[95] A driver may claim to be traveling to their mother's funeral, but not know the mother's name, or may identify a different destination than the one claimed by a passenger.[96]

Driver or passenger behavior constituting **Furtive Movements** involves motions that would suggest either a motorist's attempt to harm the officer by retrieving a weapon or the motorist merely moving to strike the officer.[97] These movements can also involve an effort to quickly conceal or destroy evidence before the officer can see or seize it.[98] The fact that the motorist appears to be under the influence of drugs has been deemed not only to provide a basis to investigate the crime of impaired driving, but also the possession of unlawful drugs.[99] Descriptions suggesting drug use, or past drug use, were annotated as **Motorist Appearance Related to Drug Use**.

Categories were consolidated to create a single generic description that encompassed specific suspicious examples identified by courts. For example, courts frequently assume that the presence of very strong air fresheners or loose baby wipes in a car is an indicator of an effort to conceal the odor of drugs.[100] Some courts have even considered cigar or cigarette smoking as an indicator of an attempt to conceal odor.[101] Rather than separately describe every possible method to cover up a suspicious

95. *See, e.g.*, United States v. Santos, 403 F.3d 1120, 1131 (10th Cir. 2005) (inability to remember names of family members, or the ages of their children, to be visited on trip); United States v. Pack, 612 F.3d 341, 353 (5th Cir. 2010) (finding reasonable suspicion, arising from conflicting answers between the passenger and driver about their relationship to the third party who had rented the car, as well as contradictions between the motorists' claim that the car was rented in Houston and the rental paperwork indicating it was rented in Pensacola, Florida.).

96. *See* United States v. Foley, 206 F.3d 802, 804-05 (8th Cir. 2000) (describing a situation where a driver claimed that the reason for the trip was a funeral, but the passenger claimed that they had not gone to a funeral).

97. If the officer believes that the furtive gesture was an effort to harm the officer by obtaining a weapon, this will justify a limited search for weapons irrespective of any effect it might have on the officer's ability to detain the car for a drug dog. *See* Commonwealth v. Buchert, 68 A.3d 911, 913 (Pa. Super. Ct. 2018). Reasonable suspicion to believe a weapon is present in the vehicle justifies an intrusion limited to those areas in which a weapon accessible to the motorist might be reached. Michigan v. Long, 463 U.S. 1032, 1049-50 (1983).

98. Pier v. State, 2018 WY 79, ¶ 24, 421 P.3d 565, 572 (Wyo. 2018) ("[Defendant] was attempting to conceal a black bag . . . [by] holding his leg in an unnatural manner.").

99. United States v. Donnelly, 475 F.3d 946, 952-53 (8th Cir. 2007) (concluding that there was reasonable suspicion to believe the motorist was using and transporting drugs because of his glassy and bloodshot eyes combined with no indication of alcohol use).

100. *See, e.g.*, United States v. Rivera, 595 F.2d 1095, 1099 (5th Cir. 1979) (observing that drug traffickers frequently use talcum powder to disguise the smell of marijuana).

101. State v. Brumfield, 42 F.3d 706, 710 (Idaho Ct. App. 2001) (puffing "excessively" on a cigar may be to mask odor of drugs).

odor, a description of any such method was annotated as a **Masking Agent**.

Similarly, courts often observed that certain types of unkempt cars raise the level of suspicion, possibly demonstrating travel so constant that time was not even taken to tidy up.[102] Courts frequently remark that empty food and drink containers, for an example, legitimately add to an officer's suspicion.[103] Specifically, when drivers have energy drinks, courts note this not only suggests a lack of tidiness, but an interest in staying awake to drive as long as possible.[104] **Indicia of Hard Travel** captures facts revealing this sort of messiness as a legitimate suspicious factor.

Officers frequently encounter more direct evidence of drug possession. Sometimes they smell the drugs themselves or identify paraphernalia, such as rolling papers or hypodermic needles.[105] When such observations are offered to a court, they are annotated as **Vehicle Contents Suggest Drugs**. Courts often conclude that the presence of some drugs suggest that more drugs are present.[106] If an officer sees, for instance, a single joint of marijuana or marijuana flakes, that discovery would fit into this category.[107]

The interest in a manageable model led to the creation of categories that include a range of possible facts, some of which, it may later be discovered, are regarded by courts to contribute more to a finding of reasonable suspicion than others. A variety of concerns about a driver's license were placed in one such category, **Motorist License**. Courts have regarded it appropriate for an officer evaluating reasonable suspicion of drug trafficking to consider a motorist's failure to possess a license, to be

102. *See* State v. Randall, 496 P.3d 844, 850-51 (Idaho 2021) (stating that the "lived-in" look of car, though innocent in itself, is an appropriate factor in considering whether there is reasonable suspicion to believe drugs are present). *But see* United States v. Townsend, 305 F.3d 537, 545 (6th Cir. 2002) (stating that given the motorists had plausible explanation for condition of the car, neither messiness nor empty food wrappers were an appropriate consideration in the reasonable suspicion analysis).

103. *See* United States v. Lebrun, 261 F.3d 731, 733 (8th Cir. 2001) (food wrappers); State v. Deviley, 2011 ND 182, ¶¶ 10-19, 803 N.W.2d 561, 566-67 (energy drink).

104. People v. Thomas, 2018 IL App (4th) 170440, ¶¶ 85-98, 115 N.E.3d 325, 340 (finding lack of reasonable suspicion observing energy drinks to be appropriate part of reasonable suspicion analysis even though "innocent drivers likewise consume energy drinks and junk food to stay awake on the road").

105. United States v. Hanlon, 401 F.3d 926, 929 (8th Cir. 2005) (rolling papers); State v. Cash, 2020 WL 1482413 (hypodermic needle).

106. *See, e.g.*, Wyoming v. Houghton, 526 U.S. 295 (1999) (finding that search of the car was justified by the discovery of hypodermic needle that driver admitted he used to take drugs). New state laws permitting medicinal and even recreational possession of marijuana seemingly will require a modification to this unstated presumption as some legal marijuana seemingly should have nothing to say about the existence of illegal marijuana (or other drug for that matter). Some courts are challenging the inference that some drugs suggest the presence of more drugs. *See* Commonwealth v. Scott, 210 A.3d 359 (Pa. Super. Ct. 2019).

107. State v. Tetreault, 2017 VT 119, ¶ 36, 206 Vt. 366, 379-80, 181 A.3d 505, 515 (Vt. 2017) (finding that marijuana flakes are a legitimate part of reasonable suspicion analysis and may themselves constitute probable cause for a search).

driving on an expired or suspended license, or to possess a temporary driver's license.[108]

Officers report a range of concerns relating to **Vehicle License Plates or Registration**. The absence of a license plate, or an expired or invalid plate or registration, would fit into this category.[109] Far less serious concerns are also described by this category. Officers frequently find—and courts agree—that out-of-state license plates add something to the reasonable suspicion analysis.[110]

Likewise, there are a number of different ways courts find driving a car belonging to another to add to the quantum of suspicion. This is true whether the car is a rental, or simply owned by someone other than the driver or one of the passengers.[111] In the interest of creating a workable model, consolidation of categories was required. It was assumed that courts attribute an equal quantum of suspicion to driving the car of a relative, or friend, who is not present as to driving a rental car. It is certainly possible that courts deem driving the rental car that a third person rented to be more suspicious than either driving a rental car that someone presently in the car rented, or driving a car owned by another.[112] Courts rarely provide guidance on the relative weight given to factors, but our categories are a first step toward creating a model with sufficient detail to be accurate, and simultaneously a small enough number of factors to be workable. This particular category regarding car ownership, however, does not allow us to investigate varying degrees of suspicion associated with the variety of ways a third party's car could be driven. Any of the three circumstances was annotated as **Unusual Vehicle Ownership**.

---

108. United States v. Olivera-Mendez, 484 F.3d 505, 510 (8th Cir. 2007) (finding that a temporary license is an appropriate factor in assessing reasonable suspicion); State v. Lawler, 2020-Ohio-549, ¶¶ 26-33, 152 N.E.3d 962, 973-74 (finding that the discovery of a motorist's suspended driver's license is a factor that may be considered in determining whether an officer can extend the stop); State v. Short, 2002-66 (La. App. 4 Cir. 01/22/03) 839 So.2d 173, 175 (deciding that a driver not having a license is a legitimate factor contributing to justification for continued detention for drug dog); United States v. Green, 52 F.3d 194, 199 (8th Cir. 1995) ("[L]ack of identification does not automatically create a reasonable articulable suspicion of criminal activity, but rather is only a factor to be considered.").

109. United States v. Garcia, No. 87-1388, 1988 WL 114167, at *1 n.2 (9th Cir. 1988) (unpublished table decision) (observing that drug smugglers often buy cars and fail to register them, making expired plates a factor adding to the analysis of reasonable suspicion).

110. United States v. Davis, 430 F.3d 345, 355 (6th Cir. 2005) (finding that driver with Michigan plates in Indiana is an appropriate factor to be considered in the reasonable suspicion analysis, especially given prior information about defendant's connections with others who were believed to be transporting drugs from Detroit to Chicago).

111. United States v. Gomez, 444 F. Supp. 3d 739, 744 (M.D. La. 2020) ("[U]se of a rental car . . . [is] consistent with the conduct of a drug courier."); State v. Kelly, 361 P.3d 1280, 1287-88 (Idaho App. 2015) (vehicle owned by a third party).

112. *See* United States v. Winters, 782 F.3d 289, 298 (6th Cir. 2015) (considering the fact that vehicle was rented by a person not present to be a factor appropriately assessing reasonable suspicion).

Since the earliest days of highway drug interdiction efforts, officers have looked for methods of communication that would set drug traffickers apart from innocent motorists. In the late 1980s and early 1990s, beepers were rare and thus raised suspicion.[113] At one point, cell phones themselves were even an indicator of possible criminal activity,[114] but today, virtually everyone has one. Possessing multiple cell phones, however, remains unusual and adds to the degree of suspicion.[115] Drug traffickers often use a difficult to trace device—frequently referred to as a burner phone.[116] As it is rare for most individuals to have more than one cell phone, there is some basis to conclude that any additional phone is a phone intended for use in drug trafficking. To make use of the large number of drug interdiction cases, possessing devices in a way that is unusual for the time period has been annotated as **Suspicious Communications Device**. Our system has achieved good results in this category, but future efforts to fine tune this category might involve some effort to associate the device, or number of devices, with the year of the decision.

It is assumed that traffickers transport their drugs out of sight. Signs of hidden compartments in automobiles are therefore seen as a good indicator of reasonable suspicion.[117] Far more frequently, however, officers identify ways of transporting items that seem unusual. For example, as innocuous as it seems, placing luggage in the back seat of a car has been seen as an effort to leave more room in the trunk for drugs, and is therefore a possible indicator of trafficking.[118] Facts suggesting such a claim are annotated as **Suspicious Storage**.

113. Derricott v. State, 611 A.2d 592, 594 (Md. App. Ct. 1992) (carrying a beeper contributed to suspicion).

114. United States v. Garcia, 52 F. Supp. 2d 1239, 1250 (D. Kan. 1999) (finding that the presence of cell phone in car to be a factor in assessing reasonable suspicion).

115. United States v. Vaughn, 700 F.3d 705, 711-12 (4th Cir. 2012) ("[The officer] had noticed four cellular phones in the center console of [the defendant's] vehicle, at least two of where were pre-paid phones . . .[which] 'are typical . . . with people involved with drugs' because no identification need be provided to obtain such phones."); United States v. Stepp, 680 F.3d 651, 666 (6th Cir. 2012) (finding that three cellular phones, one of which was pre-paid, contributes to reasonable suspicion); United States v. Townsend, 305 F.3d 537, 544 (6th Cir. 2002) (finding that three cellular phones in a car contributes to reasonable suspicion but these facts are a weak indicator).

116. United States v. Bentley, 795 F.3d 630, 638 (7th Cir. 2015) ("[D]rug dealers carry multiple phones, particularly prepaid phones that cannot be traced.").

117. United States v. Govea-Solorio, 139 Fed. App'x. 33, 35 (10th Cir. 2005) ("[I]ndications the car's interior had recently been altered to accommodate a secret compartment [contributed to reasonable suspicion].").

118. United States v. Briasco, 640 F.3d 857, 860 (8th Cir. 2011) ("[T]ravelers claimed the trunk was empty even though there was luggage on the back seat and the car's rear end was sagging."). Remarkably, though, courts have also concluded that the absence of luggage is suspicious. State v. Provet, 706 S.E.2d 513, 519 (S.C. 2011) (finding that the absence of luggage on a two-day trip contributes to reasonable suspicion).

Legal Indications of Drug Use includes all information known to law enforcement acquired by means other than the current traffic stop that suggests past or present drug possession or trafficking. If anyone in the car was previously convicted of, or investigated for, a drug crime, it would fit in this category.[119] Additionally, this category includes tips to law enforcement about specific motorists.

It is not surprising for perfectly innocent people to be apprehensive about being detained by law enforcement. Nevertheless, courts frequently consider nervousness, but with the caveat that it is given little weight in the assessment of reasonable suspicion.[120] A number of the cases the team read prior to creating the list seemed to offer a clear distinction between observations that would fit in **Physical Appearance of Nervousness**,[121] such as sweating, shaking, or twitching, and **Nervous Behavior**, which would include fast talking or pacing.[122] Annotators identified many sentences, and portions of sentences, which could describe either physical manifestation or conduct—stuttering, for instance, could be characterized as either. A future round of experiments will consolidate these two categories.

Three factors will appear odd—or even inappropriate—in this list: **Refused Consent**, **Expensive Vehicle**, and **Driver Status**. Officers and courts frequently observe that an officer's request to consent to a search was refused, but the law is very clear that this may not be used to demonstrate suspicion.[123] This category is identified for the purpose of assessing whether courts are, unwittingly, including refusal to consent in their assessments, as it appears that sometimes officers regard refusal to

119. Pier v. State, 2018 WY 79, ¶ 24, 421 A.3d 565, 572 (finding that a prior conviction for distributing methamphetamine contributed to reasonable suspicion).

120. *See, e.g.*, State v. McGinnis, 608 N.W.2d 605, 611 (Neb. Ct. App. 2000) (collecting cases on nervousness); United States v. McRae, 81 F.3d 1528, 1534 n.4 (10th Cir. 1996) ("We have held that nervousness alone is not sufficient to justify further detention . . ."); State v. Beckman, 305 P.3d 912, 918 (Nev. 2013) ("Factors such as nervousness are part of a reasonable suspicion analysis but, standing alone, carry little weight because many citizens become nervous during a traffic stop, even when they have nothing to hide."). Occasionally, however, courts will find nervousness alone, or nervousness with little more, to be sufficient for reasonable suspicion. *See* State v. McPherson, 892 So.2d 448, 452-53 (Ala. Crim. App. 2004) (holding that extreme nervousness in addition to fact that motorist told officer that he possessed a gun was sufficient for reasonable suspicion to detain the motorist for a drug dog).

121. *See* State v. Smith, 373 S.W.3d 502, 505 (Mo. Ct. App. 2012) ("[Officer] was struck by Appellant's nervousness, sweating, and shaking.").

122. *See* Commonwealth v. Thomas, 478 S.E.2d 715, 722 (Va. Ct. App. 1996) ("[Officer] observed defendant's nervous behavior, including locking the car door, pacing, and becoming excited and agitated.").

123. Florida v. Bostick, 501 U.S. 429, 437 (1991) ("[A] refusal to cooperate, without more, does not furnish the minimal level of objective justification needed for a detention or seizure."). *But see* Wade v. State, 422 S.W.3d 661, 674-75 (Tex. Crim. App. 2013) (declining "to hold that a citizen's questions or refusal to cooperate with a police request during a consensual encounter can never be a factor in determining whether an investigative stop was justified").

consent as a factor supporting a continued detention.[124]

Though not expressly illegal, **Expensive Vehicle** and **Driver Status** feel problematic. Courts have used a driver's status as male, young, or being from another state as a factor bearing on suspicion—something that was sometimes deemed more suspicious when an expensive car was being driven. In the early days of Operation Pipeline, there was widespread concern about officers profiling minority motorists in expensive vehicles.[125] This factor is still identified but in less problematic circumstances, with courts occasionally observing, for instance, that it is unusual to rent an expensive vehicle given the occupants stated nature of travel[126]. To the extent, however, courts are still using this factor in ways that appear to permit a particular type of profiling—and a review of several hundred cases reveals such cases wane after the 1990's—this case law still constitutes a statement of the law that the prosecution could use to support denial of a motion to suppress and therefore should be represented in the computational assessment.

Annotating the **Expensive Vehicle** factor so that it can be automatically identified from the approximately 40,000 judicial opinions, however, serves another important and very different goal. If this factor is playing a substantial role in reasonable suspicion assessments, identifying the extent to which this is true provides a basis for raising this concern to legislatures or courts with the power to discontinue its consideration in the suspicion analysis.

Finally, our annotation process included an **Other** category to capture categories of suspicion that were missed in the creation of the initial list of factors. As student annotators examine subsequent cases, it may be that additional factors must be included in the model, and this **Other** category allows for a recording of factors that have not already been accounted for. A few cases observe that the intentional displaying religious symbols is suspicious as an attempt to deflect an officer's suspicion.[127] At present, it does not appear that this factor appears frequently enough to justify its own category, but this is an example of a category whose significance may be realized later.

---

124. United States v. Wendfeldt, 58 F. Supp.3d 1124, 1132 (D. Nev. 2014) ("The Court is particularly troubled that Wendfeldt's refusal to consent appears to be a factor in Trooper Lee's decision to conduct the dog sniff.").

125. *See* Derricott v. State, 611 A.2d 592, 594 (Md. App. Ct. 1992) (finding that "young black males wearing expensive jewelry" while "driving expensive cars" contributed to reasonable suspicion).

126. *See* United States v. Walton, 827 F.3d 682, 685 (7th Cir. 2016) ("Officer McVicker testified that the Suburban rental caught his attention because the two did not have a need to rent such a large and expensive vehicle; there were only two occupants and one bag of luggage, yet the car cost nearly $1,000 and seated seven to eight passengers.")

127. *See* United States v. Townsend, 305 F.3d 537, 544 (6th Cir. 2002); United States v. Ramon, 86 F. Supp. 2d 665, 675 (W.D. Tex. 2000).

At present, our model assumes that courts give equal magnitude to each possible observation that fits within a single category. This may or may not accurately reflect what courts are doing with these factors, and our subsequent experiments will assess that question. For instance, if a car is stopped in Richmond, Virginia and the driver identifies the destination as Atlanta while the passenger says it's Miami, courts would regard these as inconsistent statements that contribute to a finding of reasonable suspicion. A court should, however, regard it to be more inconsistent—and thus more suspicious—if the driver said Atlanta and the passenger said somewhere far from there, like Seattle. The first set of inconsistent answers may raise suspicious but reflect nothing more than the driver's account of an intermediate stop and the passenger's identification of the ultimate destination. On the other hand, it is far more difficult to conclude that the second set of inconsistent answers is something other than a poorly coordinated attempt to hide true travel plans from the officer. Nevertheless, our factors currently treat these two set of inconsistent answers with equal weight.

Similarly, the current factors would give the same magnitude to an officer's identification of a single Febreze air freshener as it would to an officer's report that, upon approaching the vehicle, seven evergreen air fresheners were observed hanging from the rear-view mirror, twenty boxes of baby wipes strewn throughout the vehicle, and all four occupants of the car feverishly smoking strong cigars.

In short, courts may or may not regard suspicious facts as equally contributing to reasonable suspicion merely because they fit into the same category of suspicion.[128] The ultimate version of the pipeline to assess suspicion may need to find some method of accounting for the varying weights to be assigned to the different circumstances within each category. That work is beyond the scope of the current experiments, but it is assumed that an automated system that distinguishes between aggravated and ordinary examples within each category will reflect the varying values courts assign to the particular circumstances.

Models may well be developed to account for degrees of suspicion in subsequent iterations of the pipeline to improve accuracy. Of course, it is not yet entirely clear how the courts are regarding factors that exist in different degrees. Using the examples above, four air fresheners in a car likely would (and should) be regarded as more suspicious than one, but it does not seem that fourteen would be any more suspicious than four. A single air freshener is not out of the ordinary although it may be part of an effort to conceal odors, but anything more than one can be harder to

128. *See, e.g.*, United States v. Simpson, 609 F.3d 1140, 1147 (10th Cir. 2010) (observing that nervousness can carry varying weights in the analysis depending on whether the officer observes very specific characteristics suggesting extreme nervousness).

innocently explain. Similarly, if the driver describes a journey from Richmond to Atlanta, a passenger's identification of Anchorage as the final destination is more inconsistent than Seattle, but both equally reveal a poor effort to conceal the true destination. In later work on this project, an effort will be made to identify language distinguishing descriptions that minimally satisfy each category of suspicion from descriptions of circumstances that are more suspicious.

Finally, categories needed to be created for language representing the court's conclusion about whether reasonable suspicion was present. To prevent capturing language relating to other Fourth Amendment doctrines, this annotation identified only language addressing whether the facts rise to the level of reasonable suspicion. If a court, for instance, opined that the officer had a sufficient basis to believe drugs were present, this would be annotated as **Suspicion Found? – Yes**, just as language asserting that the officer lacked a sufficient legal basis to believe drugs were present would be annotated as **Suspicion Found? – No**.

B. Annotating a Sample of 211 Cases

To understand to what degree humans can agree on what factors are described in individual sentences, and to use annotations to train and evaluate a language model, we annotated sentences in legal opinions. A team of law students was trained to annotate the court language in 211 cases, which were selected from a larger group of cases having indicia of drug interdiction stops where reasonable suspicion was at issue. The Harvard Caselaw Access Project, which provides access to free electronic versions of judicial opinions and permits researchers to download cases in bulk for experiments like ours, retrieved this initial batch of cases.[129] The team used the search terms "canine," "reasonable suspicion," and "drug interdiction" to identify these cases. The terms "drug interdiction" and "canine" certainly do not appear in every case involving stops of the type we are considering. "Drug interdiction" is no part of the legal analysis and would appear in the cases only where the court chooses to note that the search and arrest were performed by a subdivision of a police force dedicated to interdiction work.[130] Certainly, this is not essential to the narrative of the case, and very often drug interdiction work is done by

129. *See* Connie Lenz, *Affordable Content in Legal Education*, 112 L. LIBR. J. 301, 318 (2020) (describing free access to cases).

130. *See, e.g.*, David A. Harris, *Car Wars: The Fourth Amendment's Death on the Highway*, 66 GEO. WASH. L. REV. 556 (1998) (observing that state police drug interdiction units in various states more frequently stopped minority motorists than officers in the same departments who were not assigned to the drug interdiction units).

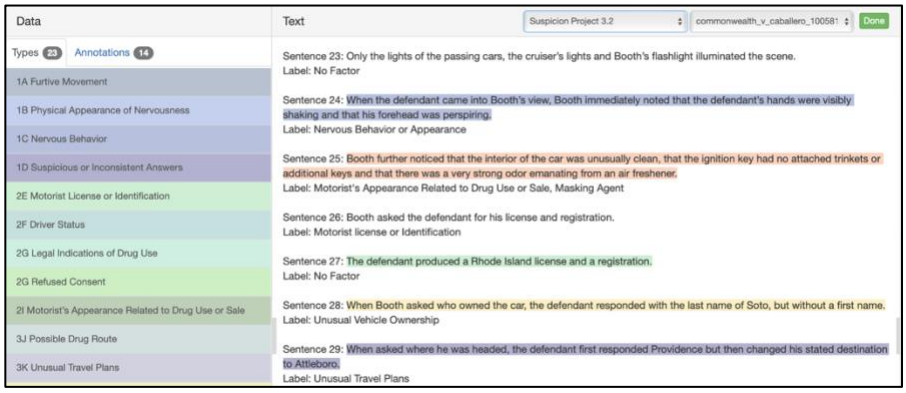
officers not specifically assigned to a subdivision with such a name. Additionally, the drug dog may not be identified as a "canine" or even mentioned at all.

Limiting the initial set of cases to those including these terms, however, decreased the likelihood of including off-topic cases. If a case uses these three terms, it is highly likely that the case will address whether reasonable suspicion exists to believe drugs were present in the vehicle. From the cases downloaded in this search, 211 were identified as likely on point and were assigned to law students to annotate for the various suspicious factors. The students were nevertheless given the option of annotating a case as **Possibly Off Point** if it appeared that this issue was not being addressed.

The 211 cases were not chosen to ensure that their results were, overall, consistent with the percentage of cases finding reasonable suspicion present and absent respectively. There is nevertheless a critical mass of cases finding the facts satisfying—and failing to satisfy—the reasonable suspicion standard. In these cases, 57% of state court decisions concluded that reasonable suspicion was present and 43% concluded that it was not. Of the federal decisions, 77% found reasonable suspicion and 23% did not.

The law students used annotation guidelines, which instructed them to identify language fitting into the categories identified above in bold. The students were told to identify sentences within the majority opinion fitting into these categories even if the sentences fit into more than one of these categories. Students annotated the cases using the Gloss annotation environment, pictured below, developed by co-author Jaromir Savelka, which allowed them to highlight text and identify its appropriate category.

Figure 1: Annotating Factors with Gloss



Two students annotated each case to ensure that the text was correctly and clearly identified as fitting within a factor's description. Differences

between the student annotators suggest one of two things. First, it is possible that the students arrived at different conclusions because the text identified did not clearly denote a particular factor. It is possible that a language model could not identify factors of legal text if trained lawyers are unable to do so. But two human annotators may arrive at different conclusions because of human error. Second, there may have simply been an incorrect identification or failure to recognize the text as fitting into any category.

The first type of inconsistency is important in understanding whether this task can be meaningfully performed with a language model. The second type of inconsistency offers no useful information and falsely suggests that the factors are less clearly identified than they actually are. To isolate the reasons for inconsistent annotations, the students were asked to compare their work. The Gloss environment made this task easy, as sentences are highlighted in various colors depending on the factor categories the annotators identify for the sentence. Where the sentences were identified as fitting into different categories, the students were asked to consider why they had arrived at differing conclusions about the text. If they disagreed about how a piece of text should be annotated, or whether it fit within any of the categories at all, they were to leave the original categories they had each assigned to the text in place. If, however, the comparison caused one of the annotators to realize that the category had been assigned in error—or that a piece of text should have been annotated but was not—then the erroneous annotation was to be corrected.

The extent of disagreement between annotators is, however, important in assessing whether the language in the opinions identifies the categories identified with sufficient consistency that a model could accurately assign language to categories. If the human annotators arrived at different answers with sufficient frequency, then the task would not be computationally possible. Using a statistical measurement called Cohen's kappa ("κ"), it was determined that there was a moderate amount of agreement among the annotators given the complexity of the task, suggesting that the classifying of the text was not beyond the ken of a language model.[131]

C. Using Language Models to Identify Factors

Above we explained the procedure by which law students annotated

131. For a full description of this analysis, *see* Morgan Gray et al., *Toward Automatically Identifying Relevant Factors*, *in* LEGAL KNOWLEDGE AND INFORMATION SYSTEMS, 53, 58 (Francesconi et al. eds., 2022).

cases. The process of annotation is costly in terms of time and money. Thus, it would be useful it was possible to identify relevant factors from cases automatically. In this Part we explain our efforts to accomplish this with language models of varying sizes and capabilities.

In our implementation, we discuss two approaches. First, we use a fine-tuned language model to learn a representation of the language that describes each particular factor and then we use that representation to identify what factor or factors a sentence is describing. Second, we use large language models ("LLMs"), particularly in the GPT family of models, to identify factors based on a detailed prompt. In simple terms, a language model works by "assigning probabilities to word sequences and predicting upcoming words."[132] The difference between a "language model" and a "large language model" generally has to do with the size of a model and its purpose. A large language model is "designed to understand and generate text like a human . . . based on the vast amount of data used to train them."[133] In other words, an LLM is a large model, trained on a vast amount of data, and can generate text. Language models are not as large as LLMs and do not always generate text.

Then we discuss our use of a language model to identify factors. We used RoBERTa, short for Robustly-optimized Bidirectional Encoding Representation for Transformers, which is capable of processing a large volume of text and assessing each word's relationship to the surrounding words.[134] To use this model to identify factors, we engaged a process called "fine-tuning", where we provided the model with sentences describing one or more factors in order for it to learn the language that describes a particular factor. For instance, consider the following. At a hearing, an officer testified that the reasons for the search were: Berry's nervousness, his uncertainly about whether his son was working or not, the fact that he was driving a rental car, the rental contract, Berry's looking down the interstate before answering some questions, Berry's failure to remember that his son lived in Decatur, a plastic garbage bag in the backseat, a the long trip from South Carolina only to stay a few hours. This sentence contains **Physical Appearance of Nervousness** (1B), **Suspicious Answers** (1D), **Unusual Vehicle Ownership** (4N), **Nervous Behavior** (1C), **Suspicious Storage** (5S), and **Unusual Travel Plans** (3K). This sentence would be provided to the model and labelled as describing these factors.

132. DANIEL JURAFSKY & JAMES H. MARTIN, SPEECH AND LANGUAGE PROCESSING 136 (3d ed. draft 2024) , https://web.stanford.edu/~jurafsky/slp3/ed3bookfeb3_2024.pdf.

133. *What Are Large Language Models (LLMs)?*, IBM, https://www.ibm.com/topics/large-language-models (last visited May 7, 2024).

134. For a description of RoBERTa, see Yinhan Liu et al., *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, ARXIV (2019), https://arxiv.org/pdf/1907.11692.pdf.

The labels assigned to each sentence by the annotation process were then used in conjunction with the RoBERTa model to allow a process based in machine learning to increasingly improve the system's ability to identify the factors in the sentences. With each experiment using RoBERTa, 60% of the 211 cases were used as training data, which the system used to learn how to correctly identify the text in each of the categories. The remaining 40% of the cases were used as test data to identify how well the algorithm produced by the training set could identify sentences containing each of the twenty-four criteria—twenty-one of which identified factors officers used to identify suspicion, two of which related to the court's judgment, and one that identified a case as being incorrectly selected for annotation. The table below reveals a typical set of results after fifteen epochs of this method.

Table 3: Quality of Factor Identification with Multi-Label Sentence Annotation

	precision	recall	f1-score	support
no_type	0.99	0.99	0.99	8381
1B Physical Appearance of Nervousness	0.92	0.89	0.90	62
2H Legal Indications of Drug Use	0.89	0.78	0.83	54
4N Unusual Vehicle Ownership	0.82	0.78	0.80	51
2G Refused Consent	0.79	0.73	0.76	30
3J Possible Drug Route	0.90	0.75	0.82	24
6W Suspicion Found? - Yes	0.92	0.83	0.88	42
5P Masking Agent	0.80	0.84	0.82	19
1C Nervous Behavior	0.78	0.85	0.81	53
3K Unusual Travel Plans	0.82	0.64	0.72	14
1D Suspicious or Inconsistent Answers	0.93	0.72	0.81	75
5Q Vehicle Contents Suggest Drugs	0.79	0.79	0.79	34
4M Vehicle License Plate or Registration	0.67	0.25	0.36	8
1A Furtive Movement	0.79	0.58	0.67	19
6V Suspicion Found? - No	0.71	0.87	0.78	31
6U Possibly Off Point	0.00	0.00	0.00	1
6T Other	0.79	0.58	0.67	19
2E Motorist License or Identification	1.00	0.91	0.95	11
2F Driver Status	0.00	0.00	0.00	0
5R Suspicious Communication Device	0.00	0.00	0.00	3
5S Suspicious Storage	0.73	0.84	0.78	19
5O Indicia of Hard Travel	0.00	0.00	0.00	6
2I Motorist's Appearance Related to Drug Use or Sale	0.92	0.85	0.88	13
4L Expensive Vehicle	0.00	0.00	0.00	3
micro avg	0.98	0.98	0.98	8972
macro avg	0.66	0.60	0.63	8972
weighted avg	0.98	0.98	0.98	8972
samples avg	0.92	0.92	0.92	8972

The precision scores in this table represent the number of times a sentence was correctly identified divided by the sum of times the sentence was correctly and incorrectly identified, or:

$$\textbf{\textit{Precision}} = \frac{\textbf{\textit{True Positive}}}{\textbf{\textit{True Positive}} + \textbf{\textit{False Positive}}}$$

The recall scores are calculated by dividing the number of times the sentence was correctly identified by the sum of the times the sentence was correctly identified and the number of times a sentence with this factor was not identified:

$$\textbf{\textit{Recall}} = \frac{\textbf{\textit{True Positive}}}{\textbf{\textit{True Positive}} + \textbf{\textit{False Negative}}}$$

The F1 scores combine precision and recall into one value; its value is twice the product of precision and recall divided by the sum of precision and recall:

$$\textbf{\textit{F1}} = 2 \text{ x } \frac{\textbf{\textit{Precision}} \text{ x } \textbf{\textit{Recall}}}{\textbf{\textit{Precision}} + \textbf{\textit{Recall}}}$$

This data reveals that, with a reasonable amount of data, RoBERTa has the capacity to identify sentences describing facts in the categories used in the annotation process. RoBERTa obtained promising precision and recall scores with a critical mass of sentences in the training data relating to a category. No categories with fifty or more sentences in the training data produced an F1 score below 0.80, and some categories achieved higher F1 scores with considerably fewer sentences. **Motorist's Appearance Relating to Drug Use or Sale** achieved an F1 score of 0.88 with thirteen sentences in the training data, and with a mere eleven sentences in the training data, **Motorist License or Identification** achieved a score of 0.95. The category **No Type** unsurprisingly had a nearly perfect precision and recall score. The overwhelming majority of sentences in cases do not fit into any of the categories designated for annotation, and most of them bear no similarity to sentences that describe facts in the categories These values are all 0 for four of the categories. RoBERTa identified sentences in the **Possibly Off Point**, **Driver Status**, **Suspicious Communication Device**, **Indicia of Hard Travel**, and **Expensive Vehicle** categories with precision, recall, and F1 scores of 0.

Although a fine-tuned RoBERTa model shows promise in identifying legally relevant factors, the process of fine-tuning is costly in many ways. First, high quality annotations need to be generated. The process of the annotation described above cost around $8,500 total and took months of work.[135] Second, it can take a long time to fine-tune a model. Given these

135. Morgan A. Gray et al., *Empirical Legal Analysis Simplified: Reducing Complexity Through Automatic Identification and Evaluation of Legally Relevant Factors*, 382 PHIL. TRANSACTIONS ROYAL

considerations, and the proliferation of LLMs, we turned to the GPT family of models to automatically identify legal factors in text. We focus on the use of GPT-4 given its performance on tasks in the legal domain such as passing the bar exam, classification of legal texts, explanation of statutory provisions, and the annotation of the roles of sentences in legal documents.[136]

Although fine-tuning can be applied to LLMs, we focus on the mechanism used to perform tasks with an LLM: prompting. A prompt is a set of instructions provided to an LLM to perform a task. To show the feasibility of factor identification with an LLM, we relied on a subset of the 211 cases identified in the RoBERTa experiment. In our case, the task is the same as it was with the RoBERTa model. However, instead of fine-tuning, we used a prompt that guides the LLM in identifying legally relevant factors in sentences. Our prompting approach relied on the annotation guidelines that were provided to student annotators to identify legally relevant sentences. Specifically, via the prompt, we provided GPT-4 with a description of the task of identifying legally relevant factors in sentences. We also described the legal problem and provided a hypothetical scenario describing the detention of a motorist on suspicion of drug trafficking. Next, we described and defined in clear detail the factors that the model would be identifying. In addition, we provided the model with specific sets of rules to follow when annotating. Lastly, we provided the model with examples showing a sentence and the ideal label or label(s).[137] The results of the factor identification are discussed in Table 4.[138] In addition to precision, recall, F1 scores, and accuracy, we also assess GPT-4's capability to "correctly identify the total number of factors present in each case. That is if a factor was present *at all* in the case, did the model identify it . . . ?"[139] We refer to this as the intersection. We also assess to what extent the model identified factors that were not present in a sentence. We refer to these as false factors.

SOC'Y A 1 (2024).

136. *Id.* at 4-5.

137. *Id.* at 10. More information about the technical details of our prompting approach can be found *supra*, note 135.

138. There are some variations between Table 3 and Table 4. In the time between the two experiments captured in these tables, some changes were made to the list of factors. Specifically, we combined factors 1B and 1C from Table 3 into a single factor **Nervous Appearance of Behavior** (1B). This was done because of the high similarity between the **Physical Appearance of Nervousness** and **Nervous Behavior**.

139. *Id.* at 12.

Table 4: Quality of Factor Identification with LLM Sentence Annotation

	precision	recall	f1-score
2E Driver Status	0.30	1.00	0.50
4K Expensive Vehicle	0.00	0.00	0.00
1A Furtive Movement	0.73	0.86	0.79
5N Indicia of Hard Travel	0.20	0.50	0.29
2G Legal Indications of Drug Use	0.75	0.92	0.82
5O Masking Agent	0.90	1.00	0.95
2D Motorist License	0.26	0.88	0.40
2H Motorist's Appearance Related to Drug Use	0.60	0.23	0.33
1B Nervous Behavior or Appearance	0.88	0.80	0.84
NF No Factor	0.98	0.89	0.94
3I Possible Drug Route	0.77	0.89	0.83
2F Refused Consent	0.66	1.00	0.80
6S Suspicion Found	0.53	0.88	0.66
6T Suspicion Not Found	0.33	0.77	0.46
5Q Suspicious Communication Device	0.44	1.00	0.62
5R Suspicious Storage	0.63	0.73	0.68
1C Suspicious or Inconsistent Answers	0.51	0.85	0.63
3J Unusual Travel Plans	0.57	0.77	0.65
4M Unusual Vehicle Ownership	0.60	0.65	0.62
5P Vehicle Contents Suggest Drugs	0.25	0.63	0.36
4L Vehicle License Plate or Registration	0.12	0.50	0.20
Macro Accuracy	0.91		
Intersection	0.97		
False Factors	2.60		

In Table 4, we can see that GPT-4 shows promising results, especially given that the model was not fine-tuned and these results were achieved with an expert crafted prompt. The resources used to obtain these results were roughly $45 and eight hours of work from a single expert. It seems possible that GPT-4 could replace the work of a fine-tuned language model on this task, thus saving time and resources.

In this Section we have demonstrated that it may be possible for LLMs to identify factors in opinions. If we were to achieve this automatically, we could generate thousands of datapoints without human annotation. In the next Section, we discuss how machine learning can be used in an interpretable and explainable way to predict case outcomes and analyze predictions. We use manually annotated factors, rather than automatically identified factors to show how our models could be useful in an ideal scenario.

III. ASSESSING EXTENDED VEHICLE DETENTIONS BASED ON REASONABLE SUSPICION WITH MACHINE LEARNING

Our approach to modelling starts with the notion that the aggregate behavior of many courts considering a factor can inform our prediction of how an individual court will evaluate a particular set of factors.[140] This may not seem like a particularly controversial premise in a common law jurisdiction. Courts, after all, tend to look to the work of other courts, even when it is merely persuasive.[141] With a totality of the circumstances test like reasonable suspicion, however, with tens of thousands of opinions in the drug interdiction context alone, courts are practically unable to seek the collective wisdom of others because that collective wisdom is so vast and unorganized.[142] The typical practice in a common law system—that courts tend to follow (or at least consider the reasoning of) other courts, even those outside their jurisdiction[143]—is debatable in this context because it is impossible for human judges to know or compile exactly what other courts have done. It is not unreasonable, however, to assume that the cases are the product of a certain common sense that has emerged that allows for predictions of how a court would address a given set of factors. Through machine learning we can leverage a similar assumption (that data about other cases can be used to inform the prediction of a particular case) and, through testing the model, assess the validity of the assumption.

140. *See* Hall & Wright, *supra* note 20, at 99. ("The insights gained from uniform content analysis of large numbers of opinions supplement the deeper understanding of individual opinions that comes from traditional interpretive techniques.").

141. *See, e.g.*, Amy J. Griffin, *Dethroning the Hierarchy of Authority*, 97 OR. L. REV. 51, 53 (2018).

142. The West Key Digest, for instance, does not catalog the factors court consider in evaluating suspicion. Such West Keys that cover the issue of prolonged detention of an automobile on the basis of reasonable suspicion include: Searches and Seizures 349, 349I, 349k23; Criminal Law 110, 110XXIV(U), 110XVII(M), 110XVII(M)18, 110k413.7, 110k413.12, 110k1181.5, 110k1181.5(3), 110k1181.5(3.1); Arrest 35, 35II, 35k60.4, 35k60.4(2), 35k63.4, 35k63.4(.5), 35k63.4(16).

143. The sense that another judge's opinion is worthy of consideration even when it is not controlling became controversial when the United States Supreme Court began to cite cases from foreign courts. When asked about this in her confirmation hearings, Justice Kagan disarmingly said that she was in favor of "good ideas coming from wherever you can get them." Mike Memoli, *Live: Elena Kagan Senate Confirmation Hearing*, L.A. TIMES, (June 29, 2010), https://www.latimes.com/archives/la-xpm-2010-jun-29-la-na-elena-kagan-hearing-live-20100629-story.html. The controversy—and her testimony—go to a bigger point about the sort of inputs that should inform a judge's analysis. If Justice Kagan merely liked the analysis of a New Zealand court, just as many justices cite to famous novelists, the reference would merely be an academic issue—the avoidance of plagiarism. *See* Scott Dodson & Ami Dodson, *Literary Justice*, 18 GREEN BAG 429 (2015) (identifying frequency with which the members of the Court in 2015 cite to literature). The fact that citation to foreign courts raises a concern reflects something about how judges are expected to reason toward conclusions. Due consideration for persuasive authority, even when it is rejected, seems to be something expected of a judge, otherwise there would have been no controversy about citations to foreign courts.

Concerns about AI application, in the criminal justice arena and elsewhere, stem in part from the fact that many people do not understand how machine learning models actually work, and that many of the models themselves do not overtly explain how predictions are made.[144] This Article tackles each of these concerns by offering a thorough explanation of the intuitively understandable models and a conceptual explanation of the more complex models used which are derived from the simpler ones. The principles behind many of the models are quite accessible, even for those without any sort of formal training in computer science or machine learning.[145] This Article offers how the explainable models process data, how the suspicious factors in our case are used, how the models arrive at the conclusion, and addresses the concerns raised about automated decision-making.

One of those concerns is bias, which this Section addresses below. Applications of AI in criminal justice have been shown to include illegitimate considerations—such as race or class—in determining outcomes. This Section examines how bias may have been unintentionally injected into our models and identifies preliminary ways to test for it. Implicit bias in a data set of judicial opinions raises philosophical questions about how to model this problem. Removing the effect of implicit bias in the data itself would be arduous at best.[146] Courts are assumed to apply precedent, and a predictive model that *changes* the opinions in the data set could misrepresent the law, and there is no guarantee any bias would be eliminated in such a process.

Nevertheless, the discovery of implicit bias in previous decisions seems to be something that decision makers with discretion ought to be aware of. We therefore propose that—to the extent implicit bias is discovered—it should be clearly identified for end users of the predictive model: judges and police departments who may elect not to find, or act on, reasonable suspicion. Alternative models that deviate from the predictions to discount the effect of bias could even be made available to users who are aware that the system is departing from the result that relied entirely on the sum of the legal authority.

Finally, we suggest that the models of reasonable suspicion will benefit from more data. In future research we will attempt to identify the

144. Garrett & Rudin, *supra* note 25, at 22-38.

145. As one commentator has observed, regression analysis is more easily interpretable and thus better satisfies the public's need for an interpretable model than machine learning. She observes that regression analysis can consider only a pre-defined and limited set of factors where machine learning can consider any possible factor, even those that come as a complete surprise to those familiar with the data. Elyounes, *supra* note 27, at 393. The machine learning models we used were limited to considering the factors we identified for the annotators who read the training cases.

146. *See, e.g.*, Vincent Malic et al. *Racial Skew in Fine-Tuned Legal AI Language* Models, 245-52 (IEEE Int'l Conf. on Data Mining Workshops, 2023).

defendant's race in a larger number of cases in our corpus, permitting an analysis of correlations between race and the categories of factors.

A. Explaining the Machine Learning Models

To assess a model's capability of predicting the outcome of a suspicion analysis (based on the factors identified in a legal opinion) and performing empirical analysis, a variety of models were tested. These are offered as proof of concept. Each test was performed on expert annotated data points, and not on the classification results from the previous Section. We employed Neural Network, Tree-Based, Linear Models, and non-parametric methods. All of these models accurately connect sets of suspicious factors with a judge's assessment of reasonable suspicion with accuracy rates exceeding 80%, with less interpretable[147] models demonstrating accuracy greater than 90%.

Ultimately, these models are predicting, based on some kind of approximation, that a court will find that a collection of factors amounts to reasonable suspicion. This is obviously not the way lawyers commonly go about assessing information. When courts explain whether reasonable suspicion is present, they do so in narrative form, connecting the particular facts with the reader's common sense view of the likelihood of guilt. In essence, this narrative description offers an assessment of how strongly a fact implicates a factor—i.e., the magnitude of the factor the court is describing. As described in the previous Section, our present work only identifies the presence of a factor. There is evidence that courts attempt to assign weight to factors to assess whether reasonable suspicion is present, however the weights assigned by judges are different from weights we calculated using machine learning models.

For example, in *United States v. Simpson*, the Tenth Circuit described how it considered the factors.[148] When the defendant, who had a prior conviction for drug trafficking, was pulled over, he was "so nervous that his whole body was shaking," nervousness that continued even after the officer told him that he was only going to issue a warning citation.[149] The defendant's travel plans were also unusual. He claimed that he had driven two and a half days from Nebraska to Reno where he spent one night

147. What makes a model "interpretable" has been characterized as ill-defined. Zachary C. Lipton, *The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability Is Both Important and Slippery*, 16 QUEUE 1, 3-4 (2016). In hopes of capturing the thrust of Lipton's paper we refer to an interpretable model as one that can be reasonably understood by humans in terms of how the model works to produce a prediction. *See id.*

148. 609 F.3d 1140 (10th Cir. 2010).

149. *Id.* at 1145.

gambling at a friend's house and was on his way back to Nebraska when the officer stopped him.[150]

In its reasoning, the court offered that, "in conjunction with other factors, criminal history *contributes powerfully to the reasonable suspicion calculus*,"[151] but observed that "nervousness is of *limited significance in determining whether reasonable suspicion exists*."[152] Though the defendant in *Simpson* exhibited extreme nervousness, which the court found worthy of more weight than ordinary nervousness, it was "somewhat reluctant to give substantial weight to that factor here."[153]

The court then offered that unusual travel plans were to be considered, but, unlike with the other two factors, gave no suggestion about the factor's strength.[154] The court did note, however, that this factor should only be considered if it is very strong.[155] The Tenth Circuit "has been reluctant to deem travel plans implausible—and hence a factor supporting reasonable suspicion—where the plan is simply unusual or strange because it indicates a choice that a typical person, or the officer, would not make."[156] The court found that a driving trip this long, with such a short stay, was sufficiently unusual and should be counted as suspicious.[157]

If other courts were to adopt *Simpson*'s view on all factors—a conclusion that certainly cannot be drawn in the absence of data—then those courts would not be considering weak factors in the reasonable suspicion analysis. If this line describes the decision making of courts, however, a model of less complexity would be required—only the existence or absence of a factor would need to be considered, not the strength of the factor as discussed by the court. At the end of the day, of course, courts are not attempting to put these factors into purely mathematical terms. *Simpson*'s analysis ultimately boils down to a narrative and a conclusion.

> In essence, we are asked to decide whether a police officer who has lawfully stopped a person is allowed to continue to detain that person for a short period when that person has a criminal record of drug trafficking, is acting extremely nervous in a situation where others typically relax, and provides evasive answers that describe a fairly implausible travel plan. We must determine whether the Constitution demands that a police officer in

150. *Id.* at 1144.
151. *Id.* at 1147 (emphasis in original).
152. *Id.* (emphasis added).
153. *Id.*
154. *Id.* at 1148-49.
155. *Id.* at 1151.
156. *Id.* at 1149.
157. *Id.* at 1151.

> such a situation to [sic] cease the immediate investigation and let that person go on his way.
>
> Although a close call, we conclude that Trooper Bowles had reasonable suspicion that criminal activity was afoot, and thus, had the right to briefly detain Mr. Simpson for further investigation.[158]

Humans can abstractly assign weights to factors, such as the court in *Simpson* did with two of the factors. This process can involve subjectivity on part of the court or attorney interpreting the facts. Although the 'weight' is not the same, with machine learning we can computationally assign weights to factors that can provide some insight into how a prediction was made. The computationally assigned weight may also be useful for courts to considering when deciding whether suspicion was present in a drug interdiction stop.

To train the models, a vector was created for each case to represent the factors identified. Each vector has twenty values, one for each category annotated.[159] If the court considering the case found factor present, the vector would equate the value in the position for that factor to be one; if the factor was not present, the value in that position would be 0. The factors' locations in this vector, using the labels from Table 2, appear below:

[1B, 4N, 3J, 2G, 2H, 1C, 3K, 4M, 1D, 5P, 6T, 5Q, 5S, 2E, 1A, 5R, 2F, 5O, 2I]

If a court found that an officer had discovered **Nervous Behavior** (1C), **Unusual Travel Plans** (3K), **Unusual Vehicle Ownership** (4N), and **Masking Agent** (5P), the following would be the vector representation of the case:

[0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]

We used 80% of our data to train the models and 20% to test how well the model could predict the presence of reasonable suspicion from the feature vectors.[160] The same training and testing data were used for each model. We used a cross-validation procedure to train the models. We trained using ten-fold cross-validation, repeating that procedure three

158. *Id.* at 1153.

159. A vector is a data structure that holds a series of values. For our purpose one can think of each vector as a code representing a case by the factors that are present.

160. This procedure is performed to get an understanding of how a trained model performs on unseen data points.

times.[161]

Table 5 below shows each model's accuracy in determining the existence of reasonable suspicion. Accuracy values in the table tell us how often the model correctly identified the appropriate finding of reasonable suspicion as a quotient of the total number of cases, as seen through this formula.

$$\boldsymbol{Accuracy} = \frac{\sum(\boldsymbol{True\ Positive}) + \sum(\boldsymbol{True\ Negative})}{\sum(\boldsymbol{True\ Positive}) + \sum(\boldsymbol{True\ Negative}) + \sum(\boldsymbol{False\ Positive}) + \sum(\boldsymbol{False\ Negative})}$$

The models we used vary in interpretability and complexity but fit into the four categories of Table 5, each of which is explained below. In some cases, more complex models build upon or are extensions of less complex models. For example, the Neural Network and Logistic Regression models share many similarities. A common apprehension regarding machine learning's use in law is the belief that machine learning models are indecipherable, "black box" models. Common misgivings also arise with respect to the model's algorithm, or, perhaps, the mathematical computations the model performs. We do not attempt to provide, in perfect detail, the specifics of each of the following models, but instead to convey an intuition of how a model goes about using factors to predict outcomes and how the model might be interpreted by a lawyer, as explained above.

161. Cross-validation is a useful method to ensure that models are generalizing well and to protect against over-fitting. For discussion of cross validation, see David M. Allen, *The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction*, 16 TECHNOMETRICS 125 (1974).

Table 5: Predictions of Reasonable Suspicion from Factor Vectors

Tree-Based Models	**Accuracy**	**Most Important Factors**
Decision Tree	0.853	Unusual Vehicle Ownership, Drug Route, Inconsistent Answers
Random Forest	0.975	Suspicious Storage, Unusual Vehicle Ownership, Inconsistent Answers
XGBoost	0.951	Drug Route, Inconsistent Answers, Suspicious Storage

***k*-Nearest Neighbor Models**	**Accuracy**	**Most Important Factors**
k-Nearest Neighbor	0.830	Not Applicable
Weighted *k*-Nearest Neighbor	0.829	Not Applicable

Regression Models	**Accuracy**	**Most Important Factors**
Generalized Linear Model	0.902	Inconsistent Answers, Drug Route, Masking Agents
Elastic Net	0.902	Inconsistent Answers, Drug Route, Masking Agents

Neural Network	**Accuracy**	**Most Important Factors**
Neural Network	0.975	Unusual Vehicle Ownership, Drug Route, Masking Agents

1. Tree-Based Models

Conceptually, Tree-Based models may be the most easily understood because they can be viewed as a graph.[162] Figure 2 shows the Decision Tree learned from our feature vectors during training. This Decision Tree was used to make decisions on test data points. As noted, we split our data into two sets. The training set represents 80% of the total data we collected. This data was used to train the Decision Tree. Based on the training process (described below), the model "learned" the tree represented in Figure 2. What the model "learned" is can be explained in the context of *how* the Decision Tree model learns. The model's name provides an insight into how it works. In Figure 2, there is an assortment of green and blue rounded rectangles called nodes. The nodes are connected by dashed lines called edges. At each node a decision is made, which determines what edge the model will follow. Green nodes are

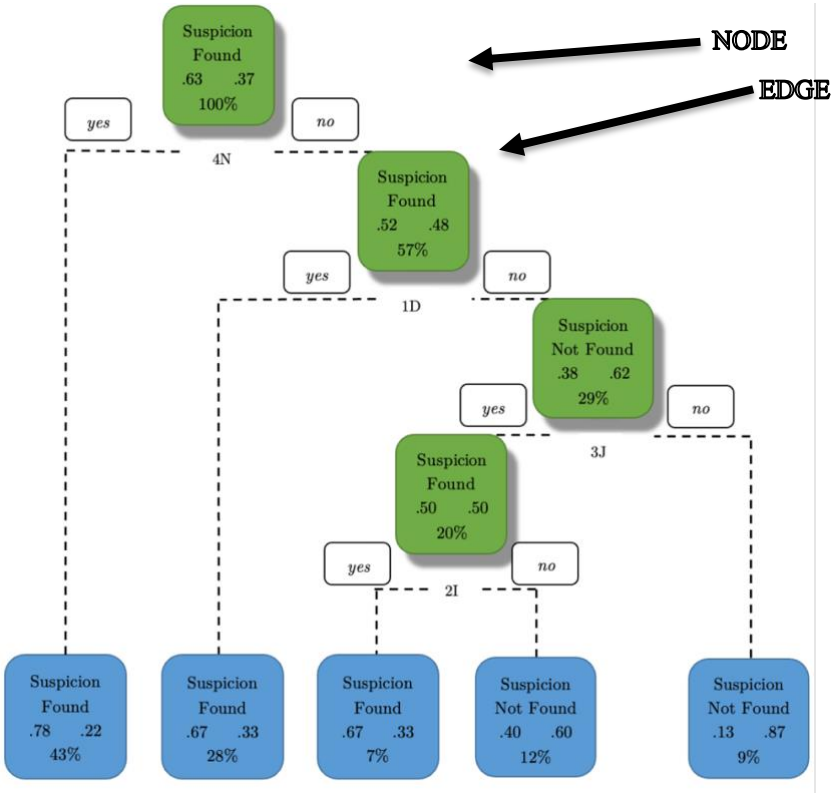
162. For an excellent and reasonably high-level description of Tree-Based models, see GARETH JAMES ET AL., AN INTRODUCTION TO STATISTICAL LEARNING WITH APPLICATION IN R 327-60 (2d ed. 2021). For a more accessible explanation, see CHRIS SMITH, DECISION TREES & RANDOM FORESTS: A VISUAL INTRODUCTION FOR BEGINNERS (2017).

known as split nodes, and blue nodes are terminal nodes. A decision at a split node tells the tree what edge to follow to the next node. This continues until the model reaches a terminal node, which represents the end of the tree, where a final classification decision is made based on the decisions made at split nodes. This kind of structure is often employed by judges to answer legal questions.[163]

In the tree represented in Figure 2, the model first assesses whether factor **Unusual Vehicle Ownership** (4N) is present. If it is, the model splits left, and predicts **Suspicion Found? – Yes** (6W) with a probability of 0.78. If not, the model then asks whether factor **Suspicious or Inconsistent Answers** (1D) is present. If it is, the model will predict **Suspicion Found? – Yes** (6W) with a probability of 0.67. The Decision Tree below concludes that if all we know is that there are no **Suspicious or Inconsistent Answers** (1D), and the court does not consider the motorist to be traveling on a **Possible Drug Route** (3J**)**, then the model will predict that reasonable suspicion is not present with a probability of 0.87. This pattern continues until all possible decisions have reached a terminal node. Importantly, the probability that the model will predict suspicion is based on the data provided in our training data. However, based on the data we have collected, this is the sequence of decisions that the tree uses to determine when to predict that suspicion is present.

163. Jonathan P. Kastellec, *The Statistical Analysis of Judicial Decisions and Legal Rules with Classification Trees*, 7 J. EMPIRICAL LEGAL STUD. 202, 205 (2010). Consider the example of determining whether negligence is present. If the judge determines that the element of duty was not met, then the analysis is over.

Figure 2: First Nodes of Decision Tree Model for Factor Vectors[164]



The diagram reveals something useful. If, for example, an officer is unable to identify inconsistent answers or that the motorist's vehicle is unusually owned, and the officer cannot claim that the motorist is on a drug route, it is much less likely that our model will predict that there was reasonable suspicion to believe drugs were present. Certainly, there are other paths to reasonable suspicion than unusual vehicle ownership, inconsistent answers, and travel on a possible drug route. Nevertheless, the empirical insights gained from this tree illuminate the important factors that lawyers and judges may want to consider when arguing or deciding cases. There are, however, important limitations that the Decision Tree does not account for.

One limitation comes from our data. Our data only considers those cases in which officers identified a basis for a search. Given that our data consists of opinions, this makes sense. No motion to suppress could be

164. Morgan Gray et al., *Automatic Identification and Empirical Analysis of Legally Relevant Factors* 108 (Proc. 19th Int'l Conf. A.I. & L., 2023), https://www.lrdc.pitt.edu/BOV/documents/Automatic%20Identification_Gray.pdf.

filed if contraband was not recovered pursuant to a prolonged detention and subsequent search. If no motion to suppress is filed, there is no reason for the court to address the issue. This is true for all models that we discuss.

A limitation of the interpretability of the Decision Tree comes from the way the tree is trained. During training, the Decision Tree looks for the feature (in this case, the factor) that reduces the mistakes (we refer to these as errors, or on the whole, error) made most by the model in determining whether reasonable suspicion is present. The first feature selected in our tree is **Unusual Vehicle Ownership** (4N). The model continues to look for features that reduce errors until some stopping criterion is met, such as the depth of the tree. The model's "depth" is defined by the number of splits. Each time the algorithm decides to split, the model goes one level deeper and considers another factor. In our case, our Decision Tree has a depth of four. Thus, the model does not reason what feature is best, but it simply decides what set of features can be used to make an accurate decision.[165] Because of this, the model does not necessarily directly consider all factors that might be present in a particular case, which a court would certainly do. However, knowing that a Decision Tree trained on these cases focuses on four factors is useful as those involved in litigating these cases may be interested to know if empirically important factors are present.

Two other Tree-Based models are useful—Random Forest and eXtreme Gradient Boosting, which build upon the classic Decision Tree. Random Forests are a staple tool in the "standard armory" of machine learning models.[166] Random Forests are an example of an ensemble model, in which a final decision is made from a group of models that are created from a data set.[167] The final model divides the data into subsets and creates a set of nodes and branches for each subset. The ultimate Decision Tree created is based on how the data was broken into nodes and branches in the majority of the subsets. To offer an analogy: assume that we have five thermometers in a pool of water. Two thermometers read that the temperature is below seventy degrees, and three read that the temperature is above seventy degrees. Based on the majority of the thermometers, we would conclude that the temperature is above seventy degrees. This is the essence of an ensemble method. Leo Brieman, who popularized the model, explains that "[a] random forest is a classifier

165. *See, e.g.*, PAUL WILMOTT, MACHINE LEARNING: AN APPLIED MATHEMATICS INTRODUCTION 127-39 (2019).

166. *See* ROBIN GENUER & JEAN-MICHEL POGGI, RANDOM FORESTS WITH R (2020); JASON BROWNLEE, ENSEMBLE LEARNING ALGORITHMS WITH PYTHON: MAKING BETTER PREDICTIONS WITH BAGGING, BOOSTING, AND STACKING (2021).

167. L. Breiman, *Random Forests*, 45 MACHINE LEARNING 5 (2001).

consisting of a collection of tree-structured classifiers . . . ."[168] "[E]ach tree casts a unit vote for the most popular class" based on the inputs.[169] Therefore, one can conceive of Random Forest—at a very high level—as a method that classifies based on the vote of an ensemble of Decision Trees.

The last of the Tree-Based models employed is another ensemble method: eXtreme Gradient Boosting, also known as XGBoost. Compared to the other models discussed thus far, this model is less interpretable. However, in simple terms, it shares some similarities to the Random Forest model in that it is based on ensembles of Decision Trees. This model, however, identifies areas in which the Decision Trees have poor performance, and then focuses attention on those areas by adding more trees. This approach makes it more challenging to conceptualize in detail what is occurring internally during the model's operation.

2. *k*-Nearest Neighbors

The *k*-Nearest Neighbor (*k*NN) classifiers are also frequently described as being very intuitive.[170] This algorithm works by using similar points, i.e., neighbors, to make a prediction. In our case, the model decides whether to predict **Suspicion Found? – Yes** or **Suspicion Found? – No** based on the similarity of a particular data point to its neighbors. There is a single parameter, *k*, that tells the algorithm how many neighbors to consider in making a prediction. If $k = 5$, the algorithm will use the five nearest neighbors to the data point for which we aim to make a prediction. The algorithm decides which neighbors are closest based on the "distance" between the points.[171] In our case, for instance, the two groups shown in red and blue represent cases in which a court found—or did not find—reasonable suspicion.

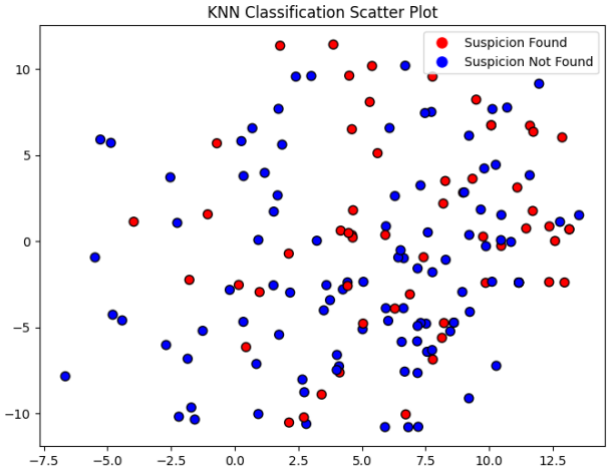
168. *Id.* at 6.

169. *Id.*

170. *See generally* WILMOTT, *supra* note 165, at 55-63.

171. An important detail discussed more thoroughly below has to do with dimensionality. In our case, we are considering twenty factors, which means that the *k*-NN model will predict the outcome based on data that is in twenty dimensions. Our predictions were made using all twenty dimensions. However, because we cannot visualize more than three dimensions, we lowered the dimensionality of the data, so it is plottable. To lower the dimensionality of our data we used logisticPCA in R. *Dimensionality Reduction for Binary Data*, GITHUB (Mar. 14, 2016), https://github.com/andland/logisticPCA.

Figure 3: *k*-Nearest Neighbors



In the figure above, we are showing predictions of **Suspicion Found? – Yes** and **Suspicion Found? – No** in a two dimensional plot that was created to aid with visualization. An image of the factor vectors for the actual model is much less intuitive and visually pleasing. First, with twenty features, one has to consider this model in twenty dimensions and how to calculate the distance between the features. To do this, we rely on Jaccard Distance. This distance metric is defined as:

$$distance = \frac{f_{01} + f_{10}}{f_{11} + f_{01} + f_{10}}$$

Consider the feature vectors that we use to represent cases such as:

Vector 1 = [1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
Vector 2 = [1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]

The distance metric counts factors that are present in one case, but not in another (f_{01} *or* f_{10}). It also measures factors that are present in both cases (f_{11}). By assessing cases in terms of what factors they share and do not share, we can assess how different the cases are. The two vectors above are very similar—there is a difference of one factor between the two. In the cases represented by *Vector 1* and *Vector 2*, the court found reasonable suspicion. The *k*-NN model predicted reasonable suspicion for both of these cases. It is possible, however, that similar cases have opposite conclusions. Nevertheless, in either scenario it is useful to know the outcomes of all similar cases.

Weighted *k*-Nearest Neighbors operate under a very similar principle.

Rather than considering the distance between all selected *k* points equally, the weighted system considers the closer points to be more significant in identifying the appropriate class than those further from the data point to be classified. In considering nearby cases represented as vectors, from this model we can infer that the cases that are the most similar will be given stronger weight in determining whether to predict suspicion found or not found. For example, if we wanted to make a prediction about *Vector 1*, and we determined that *Vector 2* was the nearest neighbor using Jaccard Distance, then *Vector 2* would have a strong influence in predicting the outcome of *Vector 1*.

A limitation on the interpretability of this model is that although the *k*-Nearest Neighbors model is easily understood conceptually, the model does not identify the role each individual factor plays in the ultimate conclusion. Thus, even though we can identify similar cases, the role of each individual factor in predicting an outcome is not known.

3. Linear Models

The Generalized Linear and Elastic Net models are both variations of classic Linear Models.[172] Figure 4, below, represents a Linear Model with an independent[173] variable on the horizontal axis and a dependent[174] variable on the vertical axis. The line shown in Figure 4 is an example of a line that could be identified by training the model. In terms of predicting an outcome, it is entirely possible, and entirely normal,[175] that no data point will lie on the line that predicts the value of the dependent variable from the independent variable. The line, however, minimizes the value of the sum of the squared distance between each data point and line, minimizing the distance that exists between the predicted value and the actual value of the dependent variable. Using common notation, the resulting line seeks to minimize the square of the difference between each

172. *See* JAMES ET AL., *supra* note 162, at 59-190.

173. We refer to independent variables with a number of different names including: inputs or features.

174. We refer to dependent variables with a number of different names including: targets, response, class, or outcome.

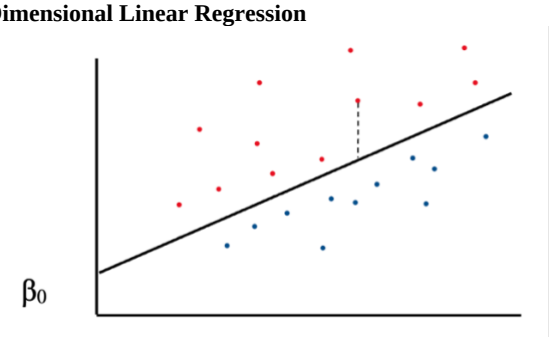
175. A model that fits training data perfectly with no error likely "overfits" the data. This means that during training, the model perfectly learned the data. One can think of this as the model having "memorized" the data. Although this may sound good at first, this is actually a bad thing. A model learned from data is just that, a model. To make good predictions, the model needs to be flexible to some degree. For example, if a law student studying for a torts exam memorized the elements of negligence only with respect to a single fact pattern and only knew of negligence within that specific fact pattern, when exposed to new facts on an exam, the student would be unable to apply the principles of negligence to the new facts. However, if the student learns the general principles of negligence, they are more likely to be able to adapt them to new cases. At bottom, the guiding principle is that a model that perfectly fits data is not useful because it cannot generalize well.

known value of the dependent variable y_i and the value that the line predicts for the dependent variable, $\hat{y}_i$.[176]

$$\min \sum_i (y_i - \hat{y}_i)^2$$

The resulting line, using a model with a single independent variable and one dependent variable, is described by the linear equation $y = \beta_0 + \beta_1 x$, where β_0 is the y-intercept of the line and β_1 is the slope. Again, because of the high dimensionality in determining whether suspicion is present, we provide a two-dimensional example for visualization.

Figure 4: Two-Dimensional Linear Regression



Linear Models can be fit using multiple independent variables, though envisioning the line requires a bit more imagination. Just like those with a single independent variable, a Linear Model on our data looks for a line that minimizes the sum of the squares of the distance between the line and each known data point. As a matter of notation, each axis is designated x_n, with n, in our case, being an integer between 1 and 20. The resulting regression line predicting values is then described by an equation very similar to the one describing a line with a single input.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

The Beta values (β), also called coefficients, represent the impact of each input (or factor, in our case) on the prediction. A large coefficient suggests that the factor plays a significant role in predicting the outcome. In a Linear Model trained on our data, factor **Unusual Vehicle Ownership** (4N) has the highest coefficient, indicating its importance in predicting suspicion. When comparing across models, **Unusual Vehicle Ownership**

176. *See* XIAN-DA ZHANG, A MATRIX ALGEBRA APPROACH TO ARTIFICIAL INTELLIGENCE 259-60 (2020).

remains crucial for both the Decision Tree and Linear Model, highlighting its strong empirical value. This factor may warrant particular attention from judges, lawyers, and police officers when assessing suspicion.

The classification task of determining whether a particular factor vector amounts to reasonable suspicion, however, requires us to take an additional step with the Linear Model. Our interest at this phase is not in determining the statistical midpoint between 0 and 1, but to assess *whether* the value is 0 or 1 (i.e., **Suspicion Found? – Yes** or **Suspicion Found? – No)**. In our case, the closer a value is to 1, the more confident the model is that the outcome is **Suspicion Found? – Yes.** If it is closer to 0, the model is more confident that the outcome is **Yes** or **Suspicion Found? – No** The regression line is thus transformed into a sigmoid function whose values tend toward the two values representing **Suspicion Found? – Yes** or **Suspicion Found? – No**.

Figure 5: Logistic Regression[177]

$$S(x) = \frac{1}{1 + e^{-x}}$$

Both of the Linear Models we employed are derived from this method. The Generalized Linear Model and Elastic Net—as forms of Logistic Regression—are very close in terms of accuracy with Random Forests, but produce results that, unlike Random Forests, are "very interpretable," as the coefficients of the resulting equation explain the impact of each factor.[178] They differ only in the sort of circumstances in the data that they correct. A Generalized Linear Model prevents errors in the prediction when the data is not evenly distributed, and an Elastic Net eliminates or reduces coefficients relating to factors that play no or little role in determining the output.[179]

177. The final step in this classification then becomes fairly simple. The equation for the $S(x)$ would produce a classifying equation from a data set that produced a regression line $y = x$. The formula for our regression, calculated above, is $y = \beta 0 + \beta 1\, x1 + \beta 2\, x2 + \ldots + \beta n\, xn$. The closer the values produced are to 0 or 1, the more certain the classification.

$$S(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

178. *See* Lin Song et al., *Random Generalized Linear Model: A Highly Accurate and Interpretable Ensemble Predictor*, 14 BMC BIOINFORMATICS 1, 1 (2013).

179. *See generally* Geoffroy Mouret et al., *Generalized Elastic Net Regression* (Joint Statistical Meeting – Section on Statistical Learning and Data Mining, 2013), https://datawisdom.ca/paper/2013-JSMProceedings-ElasticNet.pdf.

4. Neural Networks

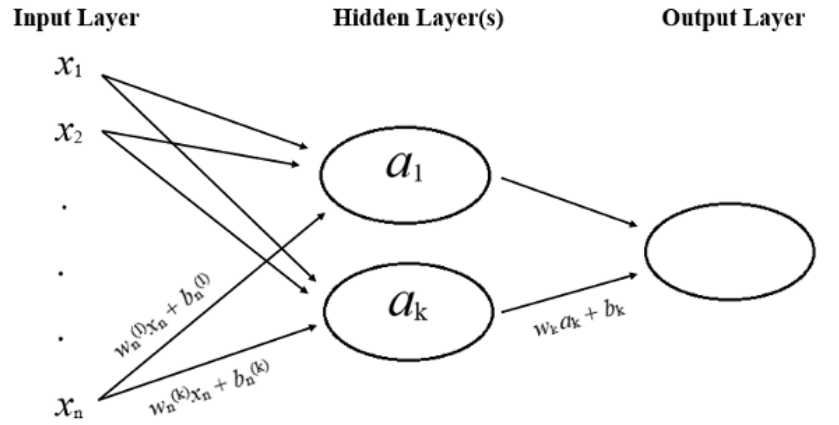
Neural Networks, at their most basic level, involve a network of Linear Models. The model takes input in an "Input Layer" and from there the model performs computation through a number of "Hidden Layers" that resemble a Linear Model. Neural Networks are often described by the number of hidden layers. Figure 6 is an illustration of a single hidden layer Neural Network. In this figure there are three layers: an input layer, a hidden layer, and an output layer. Between the input layer and the hidden layer are the input features and the associated weights. We see the weights connected to two nodes. Those nodes are called "hidden units."[180] Inside the hidden layer we see where the product of the weights and inputs are summed into each of the two hidden units. This number is then passed to an "activation" function that helps with modelling data better.

In some implementations, Neural Networks may involve many hidden layers and hidden units. Because of the complex relationships between the input, the hidden layers, and the eventual output, it can be difficult to readily interpret with exactitude what has occurred at any particular part of the network. This has led to a Neural Network being described as a black box. From a lawyer's perspective, the key takeaway is that a Neural Network's ability to learn complex relationships in data can produce accurate results, but the precise calculations of these models and their relation to the final output are not easily discernible or interpretable—they are *true* black boxes.[181]

180. An explanation of the terms "hidden units" and "hidden layer" is offered at David S. Touretzky & Dean A. Pomerleau, *What's Hidden in the Hidden Layers?*, BYTE, Aug. 1989, https://www.cs.cmu.edu/afs/cs/user/dst/www/pubs/byte-hiddenlayer-1989.pdf.

181. *See, e.g.*, IAN GOODFELLOW ET AL., DEEP LEARNING 164 (Thomas Dietrich et al. eds., 2016) (describing hidden layers).

Figure 6: Representation of Neural Network



In many circumstances, Neural Networks produce accurate outcomes. This is true in our implementation with the prediction of whether suspicion is found. However, without knowing exactly how a model relies on various factors, concerns arise as to how an output should be interpreted. A less interpretable model, such as a Neural Network could, for instance, rely on **Physical Appearance of Nervousness** as the primary factor if, contrary to a court's stated conclusions, the model actually places great weight on nervousness as a factor in the reasonable suspicion analysis. In the Decision Tree or Logistic Regression model, one would be able to examine the output of the system and identify whether it was doing so. Anyone relying on the output from a model, including courts, lawyers, or police departments, would be very interested in knowing that the model's correct predictions relied on something contrary to express judicial reasoning (such as the factor of nervousness). With a Neural Network, it is not possible to know exactly what combinations of input data it finds important, though it does appear to accurately assess reasonable suspicion.[182]

B. Minimizing Biased Decisions and Fruitless Searches

It is—and should be—insufficient to accept a model as a way to resolve legal questions simply by understanding how it works and how it processes the data it considers. It is well known that when the data relied

182. *See generally* Michael Aikenhead, *The Uses and Abuses of Neural Networks in Law*, 12 SANTA CLARA COMPUT. & HIGH TECH. L.J. 31, 55-70 (1996).

upon to train a model contains implicit biases, the output will replicate those biases. Our data is comprised entirely of judicial opinions that determine whether suspicious factors—racially neutral on their face—are present. Unlike other efforts to assist in decision-making in the criminal justice context, these models and the associated analysis try to interpretably predict an outcome of reasonable suspicion.[183] In other instances, biased data comprised of public records used in an algorithm to predict a defendant's dangerousness, or flight risk, can be attacked when it leads to more Black defendants being denied bond pending trial.[184] In our case, a model that considers the factors judges use to evaluate reasonable suspicion is less clearly subject to criticism when it relies on judicial decisions that may contain implicit biases. The models in this instance are meant to model a representation of the law itself. To modify predictions based on controlling law to account for judges' implicit biases would be to misstate the law; furthermore, a pipeline that purports to engage in a lawyer's task of summarizing persuasive authority would also be injecting a different kind of bias if the predictive values of judges' opinions were modified.[185]

183. As discussed above, others have proposed automated suspicion algorithms that consider a variety of publicly available data to create *factual* support for reasonable suspicion. *See supra* note 34 and accompanying text. This project contemplated only a system that takes as inputs just the officers' observations and assesses whether the legal standard was satisfied.

184. The bail algorithm, also known as the risk assessment algorithm or pretrial risk assessment tool, is a mathematical tool used in the criminal justice system to assess the risk of a defendant committing new crimes or failing to appear in court if released before trial. Jon Kleinberg et al., *Human Decisions and Machine* Predictions, 133 QUARTERLY J. ECON. 237, 238 (2017). It aims to provide judges with objective data to make more informed decisions regarding pretrial release and bail conditions. *Id.* The algorithm calculates a risk score based on various factors, such as the defendant's criminal history, age, current charges, and other relevant data. *Id.* Such critics of the bail algorithm purport that the tool reflects human's unconscious biases that consider factors that may not lead to later convictions. *See, e.g.*, Sean Allan Hill, *Bail Reform and the (False) Racial Promises of Algorithmic Risk Assessment*, 68 UCLA L. REV. 910 (2021); Ngozi Okldegbe, *Discredit Data*, 107 CORNELL L. REV. 2007 (2022). . Furthermore, there are issues with the lack of transparency with using AI within the criminal justice system for high stake decisions, such as bail, and the exaggeration placed on possible non-determinative factors. Doaa Abu Elyounes, *supra* note 27, at 432; John Villasenor & Virginia Foggo, *Artificial Intelligence, Due Process, and Criminal Sentencing*, 2020 MICH. ST. L. REV. 295, 353; *see generally* Chelsea Barabas, *Beyond Bias: Re-Imagining the Terms of "Ethical AI" in Criminal Law*, 12 GEO. J.L. & MOD. CRITICAL RACE PERSP. 83 (2020).

185. Our contention is very similar to Justice Jackson's recognition of the fallibility of judges in a world in which they have the ultimate authority to decide a particular issue. In *Brown v. Allen*, 344 U.S. 443, 540 (1953) (Jackson, J., concurring), he observed that, "We are not final because we are infallible, but we are infallible because we are final." The concern about modifying or selectively including judicial precedent to predict judicial outcomes is not new with algorithmic predictions. Legal treatises, which for vast periods in the history of Anglo-American law served as the primary source of law, necessarily suffered from biased inputs as their authors lacked access to all possible cases. *See* Eric J. Schwartz, *Restatement of the Law, Copyright: A Useful Resource for Practitioners,* 44 COLUM. J.L. & ARTS 425, 431 (2021) ("any well-respected treatise . . . is still a synopsis of law – entailing authors' decisions on omissions or inclusions of relevant materials and cases, and characterizations of included cases, all melded with the opinions and biases of those authors.").

There is nevertheless great value in unearthing implicit bias in judicial decisions in drug interdiction stops. Police officers and judges who may rely on the analysis capable with our pipeline have extraordinary discretion to deviate from the result that is consistent with precedent.[186] Officers do not have to detain every motorist for who they can identify reasonable suspicion.[187] Judges consider a very vague totality of the circumstances test when they evaluate reasonable suspicion.[188] For example, a judge who learned that the suspicious factors identified in a case statistically amounted to reasonable suspicion in most courts in the country could certainly consider the fact that one of the factors found—say, nervousness —was more frequently found for Black and Hispanic drivers. Such a judge would be justified in concluding that nervousness might be a function of a fear of police stops more acute in some communities and therefore nervousness should not be part of the basis for concluding suspicion exists. Police departments likewise might, with such knowledge, train their officers to rely less on such a factor with minority motorists.

The first step is to discover whether judicial opinions are implicitly biased and, if so, how. To search for bias, it is helpful to start by identifying how bias might manifest itself. Implicit bias might lead courts to find reasonable suspicion more readily in a case involving minority motorists because these cases see a lesser threshold of proof. A smaller number of factors may be sufficient, or the factors may be more readily found. If this is true, then race is effectively operating like a suspicious factor for human decisions, with race tipping close cases into the suspicious category. As our model is not designed to identify a defendant's race, one would expect it to incorrectly find reasonable suspicion present for white motorists in marginal cases and falsely conclude reasonable suspicion is absent for Black and Hispanic drivers.[189]

186. Carla R. Kock, State v. Akuba*: A Missed Opportunity to Curb Vehicle Searches of Innocent Motorists on South Dakota Highways*, 51 S.D. L. REV. 152, 175 (2006) ("Since *Schneckloth*, Supreme Court jurisprudence has shifted to give law enforcement increasing discretion to use traffic stops to meet policing priorities. . . . law enforcement has broad discretion to use traffic stops for various purposes . . . ."); Randy J. Kozel, *The Rule of Law and the Perils of Precedent*, 111 MICH. L. REV. FIRST IMPRESSIONS 37, 42 (2013) ("The decision of whether to defer to precedent would depend on the rule-of-law implications of deference in the case at hand.").

187. Amanda Charbonneau & Jack Glaser, *Suspicion and Discretion in Policing: How Laws and Policies Contribute to Inequity*, 11 U.C. IRVINE L. REV. 1327, 1333 (2021) ("The police profession is one that requires officers to use considerable judgment and discretion in the performance of their daily duties.").

188. *See, e.g.*, Lerner, *supra* note 42, at 953 (describing the vagueness of probable cause).

189. *See generally* Jon Kleinberg et al., *Inherent Trade-Offs in the Fair Determination of Risk Scores* (8th Innovations in Theoretical Comput. Sci. Conf., 2016), https://drops.dagstuhl.de/storage/00lipics/lipics-vol067-itcs2017/LIPIcs.ITCS.2017.43/LIPIcs.ITCS.2017.43.pdf.

The predictions these models make can also be skewed by race if the factors the courts rely on exist more frequently in minority populations. Two factors seem to be present more often for minority motorists. First, **Physical Appearance of Nervousness**, as some communities have sadly have experiences with police that lead to them having greater anxiety during traffic stops.[190] Courts evaluating nervousness in drug interdiction stops claim to recognize that many people are nervous when stopped by an officer and therefore only consider nervousness when it is extreme, but even then consider it only marginally.[191] If certain portions of the population are more likely to have nervousness in these stops, it seems plausible that these motorists would more frequently exhibit the form of nervousness sufficient for courts to consider.[192] Second, prior contact with law enforcement for drug possession or dealing (**Legal Indications of Drug Use**) also has the potential to be uniquely correlated with race because of the patterns of drug enforcement. These patterns result in a white drug dealer being less likely to have a criminal record than a minority drug dealer.[193] So while courts have concluded that prior interactions with drugs are highly probative of reasonable suspicion, the number of convictions, or even police encounters, for Black and Hispanic defendants may over-represent their prior drug experience compared to white defendants. Courts may determine that too much weight is therefore being given to prior illegal drug activities.

To address these bias concerns, we have begun exploring how to assess our pipeline for bias. However, this task is challenging. Because judicial opinions do not state defendants' race with any regularity, we cannot yet reliably collect this information from the text itself.[194] Nevertheless, to

190. *See* Brooks, *supra* note 87.

191. *See* United States v. Simpson, 609 F.3d 1140, 1147-48 (10th Cir. 2010).

192. *See* Paul Butler, *The White Fourth Amendment*, 43 TEX. TECH. L. REV. 245, 250-51 (2010) (observing that fear of police is quite reasonable in some communities); Illinois v. Wardlow, 528 U.S. 119, 132-35 (2000) (Stevens, J., dissenting) (same).

193. Katherine Beckett et al., *Race, Drugs, and Policing: Understanding Disparities in Drug Delivery Arrests*, 44 CRIMINOLOGY 105, 121 (2006) ("[A]lthough a majority of drug transactions involving the five serious drugs under consideration here involve a white drug dealer, 64 percent of those arrested for drug delivery in Seattle from January 1999 to April 2001 were black."); William J. Stuntz, *Race, Class, and Drugs*, 98 COLUM. L. REV. 1795, 1825 (1998) ("Because the police can more easily attack illegal street markets than other sorts of illegal markets, the crack trade has also generated more than its share of police stops and arrests. And because street markets for crack are concentrated in poor black communities, a disproportionate number of those arrests and sentences have been imposed on blacks."); *see also* David A. Sklansky, *Cocaine, Race, and Equal Protection*, 47 STAN. L. REV. 1283 (1995). In sentencing a defendant in federal court, then-Judge Nancy Gertner took into account that the defendant, who was Black, was statistically far more likely to have been convicted of minor offenses and therefore departed from the recommended sentence in light of the statistical likelihood that the defendant's prior convictions overrepresented his past criminality when compared to the criminal records of white defendants. *See generally* United States v. Leviner, 31 F. Supp.2d 23 (D. Mass. 1998).

194. Using defendants' names or surnames to identify racial bias in the texts of legal opinions has been explored in Rohan Jinturkar, *Racial Bias Trends in the Text of US Legal Opinions*. ARXIV PREPRINT

shed light on whether our methods may be biased, we have begun the pain-staking and time-consuming process of gathering this information from other sources. With knowledge of a defendant's race, we can make a number of empirical assessments about race's impact. For example, we can assess whether race is an "important" feature with respect to a model's accuracy, whether one's race is correlated with other factors of suspicion such as nervousness, whether there are more outcomes of reasonable suspicion for minority motorists than for white motorists, and other important learnings.

In subsequent phases of this research, we hope to identify the race of defendants in these drug interdiction appeals. It would hardly be surprising to learn that nervousness and prior drug encounters are more frequently found in cases involving Black and Hispanic motorists. And if found to be true, this information would be extraordinarily useful to end users of the pipeline that we discuss. Courts obviously consider these facially race-neutral factors in determining whether reasonable suspicion existed during a stop. Using precedent alone, a predictive model would inform a police officer or judge how a case is likely to be resolved in a random jurisdiction anywhere in the country. These decision-makers may, however, find it appropriate to deviate from precedent.

A totality of the circumstances test provides a judge extraordinary discretion. Additionally, police officers have complete discretion to not pursue investigations. If nervousness and prior convictions for drugs, for instance, occur more frequently when officers interact with Black and Hispanic motorists, judges could use their discretion to analyze the bias used in prior cases and to correctly conclude that nervous or prior convictions for drugs were only present in the case at bar because the case at bar involves a minority motorist—rather than assuming that those factors were present because reasonable suspicion was legally present. Thus, these judges would be able to take into consideration that certain factors are more or less predictive depending on the defendant's race.[195]

Just as producing a pipeline of analysis from the 40,000 judicial opinions of drug interdiction stops allows for the identification of implicit bias in previous decision-making, the inclusion of officers' fruitless searches in this data may reveal even more incorrect assumptions that courts have made in considering reasonable suspicion. If data indicating the suspicious factors the officer identified as well as the success or failure

ARXIV*:2307.01693* (2023); Sean Matthews et al. *Gender and Racial Stereotype Detection in Legal Opinion Word Embeddings*, 12026-33 (Proceedings of the AAAI Conf. on A.I., 2022); Douglas Rice et al. *Racial bias in legal language*. RESEARCH & POLITICS 6, no. 2 (2019): 2053168019848930.

195. Indeed, there are theories of precedent that would celebrate a modification of the method of applying the factors to preserve actual predictive value of the factors—maintaining, in other words, the spirit of precedent. *See, e.g.*, Charles L. Barzun, *Impeaching Precedent*, 80 U. CHI. L. REV. 1625 (2013); Jeremy Waldron, *Stare Decisis and the Rule of Law: A Layered Approach*, 111 MICH. L. REV. 1 (2012).

of the search was gathered, such analysis would be possible.[196]

At present, the basis for a warrantless search exists only after evidence of a crime has been discovered.[197] The predictive value of various suspicious factors can be further illuminated with data on successful and fruitless searches. With this data, it may be discovered that courts are presently giving too much weight to the fact that a rental car is being driven, or too little weight to a driver's implausible account of travel plans. It may also be that some suspicious factors innocently or coincidentally exist in certain subcommunities. Cigar smoking or heavy perfume, for instance, may be more prevalent in some populations and become part of an officer's basis for concluding there was an effort to mask the smell of drugs, even though this innocent behavior is not statistically relevant to suspicion for certain parts or subparts of the community.

Potentially, deployment of such a model in patrol cars could allow for the collection of data, permitting the creation of something previously impossible—a statistical model of suspicion. With an understanding of reasonable suspicion tied more closely to the actual likelihood that drugs are present, there would be fewer detentions for further investigation that yield no evidence of crime. This data would also demonstrate the predictive value of the factors by race. The data gathered could reveal that factors we are presently using—such as nervousness, prior drug use, or even masking agents—disproportionately fail to predict the presence of drugs for certain racial groups. As most searches in drug interdiction stops yield no evidence of crime, a burden is placed on many innocent motorists who are detained by drug interdiction units seeking to find the small number of cars actually trafficking drugs.

The models we are currently building have the potential to reduce *illegal* detentions of innocent and guilty motorists and provide the tools for decision-makers to distribute the burden of legal investigations in a more racially equitable manner. Collecting data of fruitless searches holds the potential of collecting data that could create a modified predictive model lessening the incidence of *legal* detention of innocent motors. This second step to the model could distribute the burden placed on innocent

196. A concern is frequently raised that officers will alter or shade the accounts of their observations prior to the search once drugs, or evidence of other crimes, are discovered. *See e.g.*, Christopher Slobogin, *Testilying: Police Perjury and What to Do About It*, 67 U. COLO. L. REV. 1037 (1996); United States v. Olson, 59 F. Supp.2d 725, 730 n.6 (M.D. Tenn. 1999) ("[T]he Court concludes that Trooper Ferrell's story of suspicion has, in large part, been developed between the preliminary hearings on October 31, 1998, and the time of the suppression hearing in June of 1999.").

197. *See* William J. Stuntz, *The Virtues and Vices of the Exclusionary Rule*, 20 HARV. J.L. & PUB. POL'Y 443, 447 (1997) (observing that the only beneficiaries of the exclusionary rule are those who have been accused of a crime and suppression of illegally obtained evidence often prevents, in a very visible way, the prosecution of the guilty).

motorists in a more racially equitable manner by adjusting the weight of factors that disproportionately single out minority motorists for fruitless investigations.

CONCLUSION

Reasonable suspicion—like any multi-factor or totality of the circumstances test—is unpredictable. Yet, officers conducting drug interdiction stops must apply this legal standard daily, and frequently, judges must evaluate these applications. Meanwhile, numerous judicial opinions housed online (and in law libraries across the country) sit ready to provide guidance. However, without the time or tools to access them in a meaningful way for officers performing drug stops, these opinions might as well not exist at all. Nor are they available to be used as effectively for judges and advocates as they could be.

Our preliminary experiments suggest that a system can be trained to identify the factors courts rely on in these opinions. They further suggest that, once the factors are identified with a sufficient degree of accuracy, a model could be trained to accurately predict whether a court will find these factors amount to reasonable suspicion. Police departments and private citizens alike have an interest in fewer unlawful detentions. For private citizens, there are fewer indignities that are not legally justified and for defendants, fewer violations of rights. For the police, such an improvement means less wasted time and resources and better community relations.

The use of modern technology to assist police officers has, however, failed in other contexts to enhance community trust. In part, AI is frightening because its processes are not understood—the algorithm for many is a modern oracle.[198] Hopefully, this Article has taken some of the mystery out of the methods of AI, at least in the context of identifying factors in drug interdiction cases.

Computational predictions of factor outcomes are also troubling because, for more complicated models, it is impossible to know how some decision-making models use the factors they rely upon. Particularly, in a common law system, the key to the law's legitimacy lies in its explanation

198. In modern parlance, oracles are thought of as purveyors of wisdom and knowledge. Oracle Corporation, for instance, is the third largest software company in the United States, indicating that the word has a positive connotation. CHARLES W. L. HILL ET AL., STRATEGIC MANAGEMENT: THEORY & CA: AN INTEGRATED APPROACH 293 (11th ed. 2014). Oracles in ancient times, however were (often religious) predictors of the future whose vague prophesies lacked any empirical grounding whatsoever. In one of the most famous examples of misreliance on an oracle, Croesus of Lydia was told by the Oracle of Delphi that if he attacked the Persians, he would destroy a great empire. He did so, most unsuccessfully, without knowing which empire would be destroyed. THE LANDMARK HERODOTUS: THE HISTORIES 30 (Robert B. Strassler ed., Andrea Purvis trans., 2007).

of results.[199] Our research suggests that it may be possible to achieve accurate predictions of reasonable suspicion using interpretable models.

Finally, predictive models in criminal justice are viewed with a cynical eye because the results they offer have been shown to replicate biases in the data relied upon to produce the models. The training data used to create our reasonable suspicion model cannot be shown to be bias-free—it is, after all, the product of human decision-makers. To the extent that it accurately predicts outcomes, it could, however, reduce illegal searches by relying on the very authority we expect judges to use—whether binding or merely persuasive. Nevertheless, implicit biases discovered in the model's development—biases that can be discovered and explored in this sort of data analysis—can be reported to end-users, whether police departments or judges, who have the discretion to incorporate these discoveries into their own analyses of how certain factors are to be evaluated.

Our results are preliminary, and there are reasons to be cautious. They nevertheless suggest that, for some multi-factored tests, machine learning may offer the law a way to better achieve what Justice Holmes described as the law's fundamental characteristic—predictability.[200] And in developing these models, computers may reveal biases that allow decision-makers to rethink assumptions quietly baked into precedent.

199. Mark D. Rosen, *The Structural Constitutional Principle of Republican Legitimacy*, 54 WM. & MARY L. REV. 371, 424 n.242 ("inductive reasoning . . . undergirds common law reasoning.").

200. *See* Holmes, *supra* note 12, at 460-61.