Science and Technology Law Review

Volume 23 | Number 2 Article 4

2020

Machines Finding Injustice

Hannah S. Laquer
University of California, Davis, hslaqueur@ucdavis.edu

Ryan W. Copus
University of Missouri-Kansas City, copusr@umkc.edu

Recommended Citation

Hannah S. Laquer & Ryan W. Copus, *Machines Finding Injustice*, 23 SMU SCI. & TECH. L. REV. 151 (2020)
https://scholar.smu.edu/scitech/vol23/iss2/4

Machines Finding Injustice

*Hannah S. Laqueur**
*Ryan W. Copus***

ABSTRACT

With rising caseloads, review systems are increasingly taxed, stymieing traditional methods of case screening. We propose an automated solution: predictive models of legal decisions can be used to identify and focus review resources on outlier decisions—those decisions that are most likely the product of biases, ideological extremism, unusual moods, and carelessness and thus most at odds with a court's considered, collective judgment. By using algorithms to find and focus human attention on likely injustices, adjudication systems can largely sidestep the most serious objections to the use of algorithms in the law: that algorithms can embed racial biases, deprive parties of due process, impair transparency, and lead to "technological–legal lock-in."

TABLE OF CONTENTS

* Hannah Laqueur is an Assistant Professor of Emergency Medicine at the University of California, Davis.

** Ryan Copus is an Associate Professor of Law at the University of Missouri-Kansas City.

The authors shared equally in authorship and are listed in reverse alphabetical order. We thank Jennifer Shaffer and Howard Moseley of the California Board of Parole Hearings for many patient conversations, openness to analysis, and assistance with data. We similarly thank Vanessa Nelson-Sloan and the Life Support Alliance for being a fruitful and reliable source of knowledge. We are also extremely grateful for the guidance and comments from Christopher Elmendorf, David Hoffman, Rob MacCoun, Anne Joseph O'Connell, Maya Petersen, Kevin Quinn, Sarath Sanga, and Franklin Zimring. In addition, we would like to thank our research assistants Brian Ashworth,

INTRODUCTION

The implementation of law involves the discretion of an army of decentralized decision makers, including federal and state court judges, administrative law judges, patent examiners, asylum officers, and safety inspectors. But it is well-documented that adjudicators suffer from cognitive biases, ideological blind spots, racial or gender bias, variation in moods, idiosyncrasy, and carelessness in their decision making.[1] As a result, not only are individual cases poorly decided, but the implementation of law becomes less predictable. Individuals may spend more time in prison, lose on a sexual harassment claim, or face deportation simply because they were assigned the wrong judge at the wrong time.

Historically, adjudication systems have relied heavily on secondary (i.e., appellate) review to check and monitor the discretion granted to individual decision-makers. Rather than allowing a single and perhaps cognitively biased, ideologically motivated, mood-driven, and careless decision-maker to determine a party's fate, a decision is subjected to review by a larger group of decision-makers. Then that larger group's decisions are often subject to review by a yet larger group of adjudicators. For example, federal district courts employ one judge to make an initial decision, these are reviewed by circuit courts that employ panels of three, and this decision may be subject to further review by nine judges at the Supreme Court or all of an appellate

Kevin Chen, Chance Grable-Hughes, Karalyn Lacey, Shivani Pandare, and Wenjie Xu. Finally, we thank Jonathan Simon and the Berkeley Empirical Legal Studies Fellows for their comments and feedback.

[1] A large body of research in the behavioral sciences has demonstrated a myriad of biases in human judgment that can undermine accurate or fair reasoning. *See generally* ROBYN M. DAWES, RATIONAL CHOICE IN AN UNCERTAIN WORLD: THE PSYCHOLOGY OF JUDGMENT AND DECISION MAKING (1988); RICHARD H. THALER, QUASI RATIONAL ECONOMICS (1991). Professional decisionmakers, including judges, are not immune. Indeed, a range of experimental and observational studies have demonstrated the influence of non-legal factors in judicial decisions, lending support to the legal trope that "justice is what the judge ate for breakfast." *See generally* Jeffrey J. Rachlinski & Andrew J. Wistrich, *Judging the Judiciary by the Numbers: Empirical Research on Judges*, 13 ANN. REV. L. SOC. SCI. 203, 203 (2017). Recent research suggests, for example, that the outcome of a football game, the results of the immediately preceding case, and the time of day can substantially affect legal decisions. *See* Daniel Li Chen et al., *This Morning's Breakfast, Last Night's Game: Detecting Extraneous Factors in Judging* (IAST, Working Paper No. 16-49, 2016), https://econpapers.repec.org/paper/tseiastwp/31020.htm. At a system level, studies have shown stark disparities in the rates at which adjudicators grant asylum to refugees, Jaya Ramji-Nogales et al., *Refugee Roulette: Disparities in Asylum Adjudication*, 60 STAN. L. REV. 295, 411 (2007); and provide social security disability benefits, HAROLD J. KRENT & SCOTT MORRIS, ACHIEVING GREATER CONSISTENCY IN SOCIAL SECURITY DISABILITY ADJUDICATION: AN EMPIRICAL STUDY AND SUGGESTED REFORMS 1–2 (2013).

court's judges in an en banc panel.[2] This general strategy of assigning cases to successively larger groups of decision-makers is used in almost every adjudication system.[3] The theory and hope is that larger bodies of adjudicators are more resistant to biases and whims, allowing for more accurate and consistent decisions.[4]

However, there is a major limitation to relying on secondary review, especially with the heavy caseloads that many modern adjudication systems face: we cannot afford to provide every case the full attention of an adjudication system (i.e. not every case can be reviewed en banc), and we are not very good at identifying the cases that should get review. With historically heavy caseloads in almost every adjudication system, courts must employ criteria to effectively and efficiently screen for the decisions that were most likely made in error.[5] Courts systems largely rely on an ad hoc mix of two strategies.[6] First, filing fees can leverage the litigant's judgment regarding error in the initial decision.[7] But such a willingness-to-pay approach raise concerns about inequality, and there is little indication that adjudication systems are willing to set filing fees anywhere close to the level at which they would effectively screen for meritorious appeals.[8] Second, courts can under-

[2] *Introduction To The Federal Court System*, U.S. DEP'T OF JUSTICE, https://www.justice.gov/usao/justice-101/federal-courts#:~:text=Introduction%20To%20The%20Federal%20Court%20System%201%20District,3%20Supreme%20Court%20of%20the%20United%20States.%20 (last visited Apr. 8, 2021).

[3] There are some systems that rely on a single individual for review, but even they only made the shift from multi-member review panels in response to overwhelming strains on resources. *See, e.g.*, Gerald K. Ray & Jeffrey S. Lubbers, *A Government Success Story: How Data Analysis by the Social Security Appeals Council (with a Push from the Administrative Conference of the United States) is Transforming Social Security Disability Adjudication*, 83 GEO. WASH. L. REV. 1575, 1580–81 (2015) (describing the Social Security Appeals Council's shift from *en banc* to individual review "because of the high volume of work.").

[4] *See infra* Section I.

[5] *See, e.g.*, Mike Gallagher, *Heavy Caseload, Judges' Vacancies put NM Federal Court Underwater*, ALBUQUERQUE J. (Sept. 21, 2019, 11:45 PM), https://www.abqjournal.com/1369388/heavy-caseloads-judges-vacancies-put-nm-federal-court-underwater.html; Olivia Covington, *Report: Indiana Public Defender Caseload Standards Likely too High*, THE INDIANA LAW. (July 28, 2020), https://www.theindianalawyer.com/articles/report-indiana-public-defender-caseload-standards-likely-too-high.

[6] *See* Steven Shavell, *On the Design of the Appeals Process: The Optimal Use of Discretionary Review Versus Direct Appeal*, 39 J. LEGAL STUD. 63, 63 (2010).

[7] Steven Shavell, *The Appeals Process as a Means of Error Correction*, 24 J. LEGAL STUD. 379, 384 (1995).

[8] *Id.* at 421 n.79.

take a tentative review in order to assess the likelihood of error and decide whether to subject the initial decision to the full review process.[9] But such tentative review is costly, and its preliminary search for errors can be riddled with errors of its own.

In this Article, we propose an algorithmic approach to improving review processes and legal decision-making in an ethically and technically responsible manner. Specifically, we suggest using statistical predictions of a court's decision in any given case to prioritize review of outlier decisions that are most at odds with a court's considered, collective judgment. These algorithmic predictions of decision outcomes are effectively an estimate of the percentage of judges that would disagree with a decision, were they to have decided the case. Fine-tuning the screening mechanism for appellate review means that courts, as currently structured, could spend more time correcting their own errors. Further, it presents the possibility for restructuring adjudication systems around a much more effective appellate screening procedure.

Unlike algorithms currently being proposed and deployed in contexts such as criminal justice system risk prediction,[10] the algorithmic approach we propose does not raise the same concerns of due process,[11] impaired transparency, litigant gaming, and embedded biases.

The Article proceeds in four parts. Part I discusses the challenges associated with employing algorithms to help guide decisions, focusing in particular on the controversy over the use of actuarial risk assessment instruments in the criminal justice system. Part II describes the promise and challenge of using appellate review to correct aberrant legal decisions. Part III describes the details of our general proposal to use predictive algorithms to more effectively screen cases for review.

I. THE PROBLEMS OF PREDICTIVE ALGORITHMS IN LAW

Data-driven algorithms derived through machine learning are making inroads in diverse settings ranging from business to medical diagnoses,[12] and

[9] *See id.* at 83–84.

[10] *See* Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 809 (2014).

[11] Despite advances in computation power and computer science that have improved our ability to generate predictive algorithms, these developments have not enhanced our ability to capture causes of (i.e., reasons for) an outcome. Thus, even if algorithms can correctly forecast criminal risk, they fundamentally cannot provide an account of why a defendant was deemed high (or low) risk. For more discussion, see *infra* Section II.A.

[12] *See, e.g.*, James S. Moore, *An Expert System Approach to Graduate School Admission Decisions and Academic Performance Prediction*, 26 OMEGA 659, 659 (1998) (applying machine learning methods to graduate admissions); Ziad Obermeyer & Ezekiel J Emanuel, *Predicting the Future – Big Data, Machine Learning, and Clinical Medicine*, 375 NEW ENG. J. MED. 1216 (2016) (on the

they have begun to enter the legal system primarily in the context of criminal risk prediction.[13] Jurisdictions across the country are increasingly deploying risk prediction algorithms to aid in decision tasks such as setting bail, determining sentencing, and deciding whether or not to release on inmates on parole.[14] Algorithms promise to minimize human bias and the inconsistent application of the law. At the same time, the use of risk prediction tools in the criminal justice system has been a source of considerable controversy.[15] Their growing deployment raises important technical and ethical questions that we briefly describe below.

A. Due Process, Transparency, and Litigant Gaming

A core complaint regarding the use of predictive algorithms in criminal justice decisions is that they inherently conflict with due process rights to an individualized assessment and explanation.[16] When a judge makes a decision, she can provide intelligible explanations and justifications for the decision in the particular case. This might include reasoning that because the defendant committed multiple violent crimes in the recent past, she believes he is likely to commit violence again and therefore poses a danger to society in the near future and should be incarcerated. While such justifications may or may not reliably describe a judge's internal thought process, we can at least hold judges accountable for providing some plausible justification. Predictive algorithms, on the other hand, do not, at least directly, provide such a justification. A defendant is labeled "high risk" because of a potentially complex set of statistical relationships between predictor variables and criminal offending, proxied by arrest or conviction. Inherently, a statistical model does not provide a justification for a given individual defendant. The justification for a decision guided by a predictive algorithm necessarily occurs at an aggregate and more abstract level. Because these statistical predictions can be more accurate than the predictions of judges,[17] we recommend that judges consider the risk score in making their prediction. This level of justification is one

promise of machine learning for medicine); Zhenning Xu et al., *Effects of Big Data Analytics and Traditional Marketing Analytics on New Product Success: A Knowledge Fusion Perspective*, 69 J. Bus. Res. 1562, 1564 (2016) (describing the rise of algorithms in commercial settings).

[13] *See* Starr, *supra* note 10.

[14] *Id.*

[15] *See, e.g.,* Greg Ridgeway, *The Pitfalls of Prediction*, 271 Nat'l Inst. of Just. J. 34, 35 (2013).

[16] *See, e.g.*, Anne L. Washington, *How to Argue With an Algorithm: Lessons From the Compas-Propublica Debate*, 17 Colo. Tech. L.J. 131, 140–42 (2018).

[17] *See* Zhiyuan Lin, *The Limits of Human Predictions of Recindivism*, Sci. Advance, Feb. 2020, at 1, 1.

society may or may not be willing to tolerate, and there remains ongoing scholarly debate as to whether it is incompatible with due process rights.[18]

The concern that statistical models cannot provide an individual justification because they rely on averages and correlations is heightened by the fact that predictive algorithms built with machine learning are inherently "black boxes," meaning we can see the input and the output, but what happens in between is opaque.[19] Machine learning generally affords superior predictive performance over simple regressions by better capturing complex relationships between variables, but the cost is that its formulas are much harder to interpret than simple linear regression model.[20]

The black box problem is further exacerbated by the fact that algorithms, at least as they are currently implemented, are often completely hidden from litigants and the public.[21] Many of the risk assessment instruments currently in use in the criminal justice system have been developed not by the state but by private companies who argue that the specific statistical formulas used to determine risk are proprietary, and thus shielded from review by the public.[22] Thus, the problem for due process becomes not simply that statistical models cannot provide individual-level justification, or that machine learning algorithms are difficult to explain statistically (the "black box" problem), but that the formulas used are hidden from those to whom they are applied.

Recent litigation has highlighted this unresolved legal issue. In *Loomis v. Wisconsin*, Loomis argued his due process rights were violated when the sentencing court relied on the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) criminal-risk instrument in his sentencing because proprietary nature of COMPAS prevented him from verifying or challenging the algorithm's accuracy and scientific validity and failed to pro-

[18] *See, e.g.*, Katherine Freeman, *Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in* State v. Loomis, 18 N.C. J.L. & TECH. 75,78 (2016).

[19] *See* Erin E. Kenneally, *Gatekeeping Out of the Box: Open Source Software as a Mechanism to Assess Reliability for Digital Evidence*, 6 VA. J.L. & TECH. 13, 14 (2001).

[20] *See* Eric Vardon, *Machine Learning vs. Predictive Analytics: Which is Better for Business*, FORBES (June 12, 2020, 8:20 EDT), https://www.forbes.com/sites/forbesagencycouncil/2020/06/12/machine-learning-vs-predictive-analytics-which-is-better-for-business/?sh=777a83cb4b5e (discussing the use of machine learning for business analytics).

[21] Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1254 (2008) ("The opacity of automated systems shields them from scrutiny. Citizens cannot see or debate these new rules. In turn, the transparency, accuracy, and political accountability of administrative rulemaking are lost.").

[22] *Id.* at 1290.

vide him an individualized sentence.[23] The Wisconsin Supreme Court held the trial court's use of the algorithmic risk assessment did not violate due process rights even though the algorithm used to produce the assessment was not disclosed.[24] The court reasoned that the defendant could evaluate the accuracy of the data (all publicly available) used in the algorithm's construction.[25] In 2017, the United States Supreme Court denied a petition for certiorari.[26]

Although companies have generally argued that their algorithmic formulas cannot be revealed simply because they are proprietary, there is perhaps a more serious objection to full transparency: given public access, parties might be able to adapt to models in ways that undermine model accuracy.[27] The problem of litigant adaptation comes in two forms. First, users of the legal system may be able to strategically alter their "input variables" so as to obtain more favorable algorithmic recommendations. Second, public access could result in parties entering adjudication systems that they would not have entered absent knowledge of the algorithm, which could create a disjunct between the population upon which an algorithm was built and the population to which it is applied. Consider an inmate who, absent an algorithm, would have chosen to defer his parole hearing because of his correct belief that parole commissioners would have judged him unsuitable for release. Despite the inmate's weak case for parole, he may share some characteristics with inmates who have a low recidivism score. Knowledge of his algorithmic recommendation might convince the inmate to proceed with his scheduled parole hearing. In such a case, the model would incorrectly inform commissioners that the inmate has a low probability of recidivism.

B. Embedded Bias

The second set of concerns regarding predictive algorithms center around their potential to embed biases while appearing to produce scientifically objective results. Much of the current debate centers around concerns regarding algorithms reproducing and exacerbating racial bias.[28] There are

[23] Wisconsin v. Loomis, 881 N.W.2d 749, 753 (Wis. 2016).

[24] *Id.* at 772.

[25] *Id.* at 761–62. Regarding the problem of individualized sentencing, the court acknowledged the concern that COMPAS scores only provide predictions based on group aggregations, but the problem was lessened because courts have the discretion to not follow risk score recommendations. *Id.* at 764–65.

[26] Loomis v. Wisconsin, 137 S. Ct. 2290 (2017).

[27] Michael A. Livermore, *Rule by Rules*, *in* COMPUTATIONAL LEGAL STUDIES: THE PROMISE AND CHALLENGE OF DATA-DRIVEN LEGAL RESEARCH (Ryan Whalen ed., 2020).

[28] *See, e.g.*, Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

also fundamental statistical and technical biases that can corrupt the validity of predictive algorithms. We briefly describe both the technical and normative bias concerns.

The first technical issue is the problem of "selective labels."[29] The crux of the problem is the mismatch between the dataset used to build an empirical risk prediction algorithm and the set of individuals to whom the algorithm is applied.[30] For example, consider an algorithm used to assist parole decisions. The model is fit on a dataset of individuals granted parole—the group for which there will be data on subsequent offending—but the model is applied to all parole-eligible inmates. The information about paroled individuals (that was used to develop the prediction algorithm) may not provide accurate forecasts for individuals that a parole board would not have paroled. Judges do not, presumably, release observably similar individuals randomly, so there is reason to worry about the application of forecasts of paroled inmates to the entire population of parole-eligible inmates.[31]

A second concern, and one that intersects with concerns of racial bias, is the problem of measurement. Risk assessment models in the criminal justice system aim to forecast future criminal behavior, but criminal behavior can only be measured imperfectly with administrative data on arrests or convictions. Understanding the outcome measure upon which a predictive algorithm was built, and its limitations, is critical to evaluating the validity of a predictive algorithm.[32] Predictive algorithms can only be as good as the underlying data upon which they are trained.[33] For example, an algorithm built using any arrest as the outcome, including low level violations for drug possession or loitering, may result in a risk assessment that is not accurately assessing serious criminal or violent risk, often the real interest of decision-makers. Further, it may produce artificially high-risk scores for individuals in heavily policed areas, areas that are often disproportionately black and Latino. For example, there is as evidence that suggests black men are more

[29] Himabindu Lakkaraju et al., *The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables*, PROCEEDINGS OF THE 23RD ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 275, 275 (2017).

[30] *See id.*

[31] *See id.* at 277.

[32] *See* Matt Henry, *Risk Assessment: Explained*, THE APPEAL (Mar. 25, 2019), https://theappeal.org/risk-assessment-explained/#:~:text=A%20risk%20assessment%20that%20is,commit%20crimes%20in%20the%20past.

[33] *See* Vikram Singh Bisem, *How to Ensure Data Quality for Machine Learning and AI Projects*, MEDIUM (Dec. 17, 2019), https://medium.com/vsinghbisen/how-to-ensure-data-quality-for-machine-learning-and-ai-projects-c8af1fe18c57.

likely to be caught for drug possession than their white counterparts.[34] A risk instrument built using outcomes that included arrests for drug possession will generate erroneously inflated risk scores for black males. Similarly, if black men are more frequently penalized for technical parole violations, a measure of reoffending that includes such violations of parole will unfairly increase risk scores for black men.

Beyond concerns that input or output data in an algorithm may be biased or inaccurate, some have objected to algorithms in the criminal justice system on the principle that they amount to computerized racial profiling.[35] Whether race itself, or variables that are correlated with race, are used in a model, algorithms by definition make statistical generalizations about groups that may or may not apply to any given individual. Legal scholars such as Sonia Starr have argued statistical sentencing based on gender and socioeconomic characteristics is unconstitutional: "the Supreme Court has squarely rejected statistical discrimination—use of group tendencies as a proxy for individual characteristics—as permissible justification for otherwise constitutionally forbidden discrimination."[36]

Concerns about biased algorithms have raised corresponding questions and debate as to what would constitute a "fair" algorithm. In perhaps the most widely publicized analysis of risk assessments and their potential for bias, in 2016 ProPublica reported that COMPAS sentencing recommendations were racially biased, generating more false positives for black defendants than white defendants (i.e., the algorithm incorrectly predicted that black defendants would reoffend more often than it incorrectly predicted that white defendants would reoffend).[37] In light of the ProPublica analysis, recent scholarship has formalized notions of algorithmic fairness—calibration (white and black defendants with equal scores reoffend at equal rates), predictive equality (equal false positive rates by race), and statistical parity

[34] *See, e.g.*, *A Tale of Two Countries: Racially Targeted Arrests in the Era of Marijuana Reform*, Am. Civ. Liberties Union (Apr. 17, 2020), https://www.aclu.org/news/criminal-law-reform/a-tale-of-two-countries-racially-targeted-arrests-in-the-era-of-marijuana-reform/ ("The proof is in the data: Nationwide, Black people are 3.6 times more likely than white people to be arrested for marijuana, despite similar usage rates.").

[35] *See, e.g.*, Taylor Mooney & Grace Baek, *Is Artificial Intelligence Making Racial Profiling Worse?*, CBS News (Feb. 20, 2020, 7:19 AM), https://www.cbsnews.com/news/artificial-intelligence-racial-profiling-2-0-cbsn-originals-documentary/ ("[D]uring the public comment period at a police commission meeting reviewing the audit, a community member voiced his concerns that location-based predictive policing is a covert way to justify racial profiling.").

[36] *See* Starr, *supra* note 10, at 827.

[37] Angwin et al., *supra* note 28.

(equal detention rates by race)—and shown that they cannot all be satisfied at once.[38]

Of course, when considering the potential for algorithms to be biased, it is important to consider the alternative. Human decisions can also be biased, and perhaps more so.[39] Conversations regarding the normative bias of algorithms have largely neglected the question of how algorithms compare to the alternative.[40] At the same time, biases embedded in algorithms may be particularly troubling because algorithms have the potential to be easily deployed at a large scale.[41] Furthermore, while bias in human decisions is naturally corrected as societal norms progress, algorithms pose a danger of automating and cementing historical norms, making bias resistant to societal progress.[42]

C. Outcome Models are Not Generalizable

Finally, even if the above problems with algorithms could be overcome, the algorithmically guided decision-making is not applicable to many legal domains because in most legal contexts there is not a plausible proxy for the "correct" or "good" decision upon which to develop an algorithm.[43] Unlike the criminal justice system, where risk prediction is centrally embedded in much of the decision-making, it is less clear what outcome proxy could significantly help a judge decide whether, for example, a workplace was a hostile environment, a contract was breached, or an individual is entitled to social security disability.

The proposal we detail in Part III provides a strategy for employing predictive algorithms that can be extended to all adjudication systems and largely overcomes the objections described above. Rather than algorithms being used to guide judicial decisions, this alternative use employs algorithms to find and funnel appellate resources to those decisions that are most incompatible with the court's collective judgment. More specifically, we pro-

[38] Joe Kleinberg et al., *Inherent Trade-Offs in the Fair Determination of Risk Scores* 67 Innovations in Theorhetical Comput. Sci. Conference 43:1, 43:17 (2017).

[39] *See, e.g.*, Joe Kleinberg et al., *Discrimination in the Age of Algorithms* 3 (Nat'l Bureau of Econ. Rsch., Working Paper No. 25548, 2019) Nat'l Bureau of Econ. Rsch., https://www.nber.org/papers/w25548.

[40] *See* Michael Li, *Addressing the Biases Plaguing Algorithms*, Harv. Bus. Rev. (May 13, 2019), https://hbr.org/2019/05/addressing-the-biases-plaguing-algorithms.

[41] Richard M. Re & Alicia Solow-Niederman, *Developing Artificially Intelligent Justice*, 22 Stan. Tech. L. Rev. 242, 255 (2019).

[42] Rebecca Crootof, *Cyborg Justice and the Risk of Technological-Legal Lock-In*, 119 Colum. L. Rev. F. 233, 235 (2019).

[43] *See* Leslie Green & Thomas Adams, *Legal Positivism*, Stan. Encyclopedia of Philosophy (Dec. 17, 2019), https://plato.stanford.edu/entries/legal-positivism/.

pose targeting appellate review at those decisions that deviate most from a data-driven forecast of a court's decision. But we first briefly discuss the promise and problems of appellate review.

II. THE PROBLEMS OF APPELLATE REVIEW

Foundational to most adjudication systems is a process of secondary review.[44] An initial decision, generally made by a single or small group of individuals, is reviewed by a larger group of individuals. Secondary review serves two main functions: error correction and law development.[45] The focus of this essay is on error correction, although the general framework can be extended to law development as well.[46]

On what theories does the idea that secondary review aids error correction rest? In other words, what justifies the belief that an appellate decision is more likely to remedy an initial error as opposed to simply affirming a lower-court error or, worse, introducing error where there was none before? We identify four main justifications for such a belief.

First, because appellate decisions are generally made by multi-member panels, judges are afforded the opportunity to deliberate and contribute diverse perspectives to a given decision. Deliberation can help prevent ill-considered decisions by demanding that judges provide reasoned accounts for their views and test their ideas against one another. This exchange of ideas and information can sharpen arguments and result in better judgments.[47]

Second, even without deliberation, multi-member decision-making may yield superior judgments simply as a function of the intrinsic value of judgment aggregation. The most explicit articulation of this theory is Condorcet Jury Theorem, which has been offered as way to understand the widespread belief that a multiplicity of judges will produce "better" decisions than any

[44] *See* J. Dickson Phillips Jr., *The Appellate Review Funciton: Scope of Review*, 47 L. & CONTEMPORARY PROBLEMS 1, 1 (1984).

[45] *See* Marin Levy, *Judicial Attention as a Scarce Resource: A Preliminary Defense of How Judges Allocate Time Across Cases in the Federal Courts of Appeals*, 81 GEO. WASH. L. REV. 401, 424–429 (2013); Shavell, *supra* note 7, at 379.

[46] *See* Ryan Copus, *Statistical Precedent: Allocating Judicial Attention*, 73 VAND. L. REV. 605, 628 (2020).

[47] *See, e.g.*, Stephen H. Legomsky, *Learning to Live with Unequal Justice: Asylum and the Limits to Consistency*, 60 STAN. L. REV. 413, 447 (2007); Cass R. Sunstein, *Group Judgments: Statistical Means, Deliberation, and Information Markets*, 80 N.Y.U. L. REV. 962, 963 (2005). While there are virtues to deliberation, it has also been argued that deliberation can impair the quality of decision-making insofar as group decision-making may devolve into "groupthink" in which groups will tend toward uniformity and censorship or group polarization in which camps will come to simply ignore each other. *See id.* at 965.

single individual judge.[48] The central idea of Condorcet's Jury Theorem is that a group operating under majority rule is more likely to make an accurate decision than any random member of the group deciding alone.[49]

Third, because secondary review is structured such that only a fraction of initial decisions are subject to review, courts can invest greater resources (in addition to more judges) in the review stage. For example, with lighter caseloads, courts might be able to select higher-quality judges, hire additional help (e.g., law clerks, staff attorneys, and other aides), and allow decision-makers to spend more time with each case.[50]

Finally, insofar as the appellate screening mechanism is effective, secondary review will tend to correct errors irrespective of any decision-making advantages enjoyed by an appellate tribunal. For example, if we imagine a screening mechanism that was so effective as to only allow errors into the review system, reviewers could reduce errors by simply flipping coins to make case decisions.[51] And with fewer cases to review, courts could increase the panel size to take advantage of deliberation and judgment aggregation.[52]

Currently, there are two broad approaches to determining which cases to provide with secondary review. One general approach is to rely on barriers to appeal, such that only parties who believe they have a good chance of prevailing on appeal will undertake the costs of advancing to the appellate stage.[53] Approaches of this type seek to leverage the private information litigants have about the merits of their own cases.[54] Such barriers include filing fees and indirect costs such as extensive paperwork and long delays.[55] The

[48] *See* Louis A. Kornhauser & Lawrence G. Sager, *Unpacking the Court*, 96 YALE L.J. 82, 98 n.20 (1986); Paul H. Edelman, *On Legal Interpretations of the Condorcet Jury Theorem*, 31 J. LEG. STUD. 327, 327 (2002).

[49] Franz Dietrich, *The Premises of Condorcet's Jury Theorem are not Simultaneously Justified*, 5 EPISTEME 56, 56 (2008). In the classic form, this is shown to hold true so long as each individual's probability of making the right decision is greater than fifty percent, and the group members' votes are cast independently of one another. Recent work has shown this assumption can be relaxed—one need only to assume that the average of the individuals' probabilities is greater than 0.5 for the central insight of Condorcet's Jury Theorem to apply. *See id.* at 60.

[50] Copus, *supra* note 46, at 636.

[51] Condorcet's Jury Theorem holds true so long as each reviewer's probability of being correct is greater than fifty percent. So, if only errors are reviewed, the number of erroneously-decided cases would logically decrease if at least fifty percent of the errors were corrected via coin flips. *See* Dietrich, *supra* note 49, at 60.

[52] Copus, *supra* note 46, at 647–48.

[53] *See* Shavell, *supra* note 6, at 64.

[54] *Id.* at 65.

[55] *Id.* at 69–70.

second general approach is to rely on some form of preliminary assessment by the court to determine whether an initial decision has a reasonable probability of being reversed.[56] For example, the United States Courts of Appeals hires a vast array of central staff to review cases and determine which ones are worthy of assignment to actual judges.[57] Both approaches—barriers to appeal and preliminary review—have substantial limitations.

Raising the cost of appeal raises obvious equity concerns. Further, indirect costs can involve considerable social waste.[58] Whether or not these issues are significant enough to warrant abandoning cost as a screening mechanism, adjudication systems are reluctant to fully exploit the costs of appeal as the primary screening tool. For example, filing fees in the federal and state appellate courts are almost certainly too low to encourage effective self-screening by litigants.[59]

Engaging in some form of preliminary assessment to screen for meritorious cases is time consuming and costly.[60] Additionally, preliminary assessments made by central staff are likely to be, by virtue of being preliminary, inaccurate. They are necessarily conducted under suboptimal conditions, perhaps by individuals who are not even judges, without extensive research or reflection, and with little oversight.[61]

In summary, there are serious limitation to current strategies for separating the cases that need additional review from those that do not need it. Algorithms can help.

III. PREDICTIVE ALGORITHMS FOR APPELLATE REVIEW

We propose an algorithm in which the target of the prediction is a judicial decision. The algorithm is trained on a court's historical decision outcomes and can then predict what the court's decision would be on new case data. Such an algorithm can be thought of as "synthetic" crowdsourcing: it aggregates judgments across and within decision-makers, leveraging the wisdom of the crowd and cancelling out arbitrary and contingent factors to minimize inter- and intra-judge inconsistency.[62] This noise-purified model of

[56] *Id.* at 69.

[57] Levy, *supra* note 45, at 416.

[58] *See* Shavell, *supra* note 6, at 92–93.

[59] *See id.* at 95; *see also* Shavell, *supra* note 7, at 421 n.79.

[60] *See* Shavell, *supra* note 6, at 69–70.

[61] For example, appeals by pro se litigants are frequently assessed by staff attorneys rather than Article III judges. Richard Posner, Reforming the Federal Judiciary: My Former Court Needs to Overhaul its Staff Attorney Program and Begin Televising its Oral Arguments 8 (2017).

[62] We present the case for the guiding dichotomous outcomes, but the framework could be straightforwardly extended to continuous outcomes. In fact, continuous outcomes provide a particularly promising context, as they provide a more

decision-making produces what is effectively the collective judgment of a court,[63] and we argue that it can be used to identify decisions that are most incompatible with the court's general jurisprudence.

In what follows, we elaborate on the core intuition that the algorithm can be understood as simulating a world in which each judge casts multiple independent votes in every case, explain how to best pursue the simulation effort, and suggest how the results from the simulation can improve decision-making systems in a technically and ethically responsible manner.

A. The Simulation Goal

As described in the preceding section, a core goal of appellate review is to correct errors.[64] At the same time, with rise of legal realism has come the consensus that there is rarely an objectively correct concept of "error."[65] We thus adopt a non-substantive and realist view of error: a decision is made in error if most judges would think it is. Of course, it is possible that, even if all judges would believe a particular decision was made in error, someone else might vigorously argue that it is correct. Arguments in favor of under-represented conceptions of error are no doubt important to the progress of justice. But the appellate system is structured so as to enforce the consensus view—secondary review allows judges to address decisions that, in the view of a broader set of judges, are wrong. Similarly, we take as our goal not to advance our own idiosyncratic views of justice but to help courts advance

accurate measurement of judicial judgment. With dichotomous outcomes, a one or a zero (e.g., a grant or deny) does not let us know how confident the judge's assessment is. For example, we don't know whether a grant was just barely a grant or whether it was a decisive grant.

[63] Our approach is an extension of judgmental bootstrapping, an idea conceived in the early 1900s in relation to predicting corn crop quality and later developed in a variety of fields in the 1960s. In a 1971 review of the research, Robyn Dawes coined the term "bootstrapping." HANS G. DAELLENBACH & ROBERT L. FLOOD, THE INFORMED STUDENT GUIDE TO MANAGEMENT SCIENCE 158 (2002). The central insight is that the fitted values from a regression model of expert judgments, by eliminating the uninformative variance or noise, will often correlate higher with predicting the outcome variable than the actual expert judgments themselves.

[64] Levy, *supra* note 45, at 424–25.

[65] Joseph William Singer, *Legal Realism Now*, 76 CAL. L. REV. 465, 467 (1988) ("We are all realists now . . . All major current schools of thought are, in significant ways, products of legal realism. To some extent, we are all realists now."); *see also* Gregory S. Alexander, *Comparing the Two Legal Realisms—American and Scandinavian*, 50 AM. J. COMP. L. 131, 131 (2002) ("'We are all Realists now,' as the saying goes."); *see also* Brian Leiter, *Rethinking Legal Realism: Toward a Naturalized Jurisprudence*, 76 TEX. L. REV. 267, 267 (1997) (beginning the essay with the "cliché" that "we are all legal realists now").

their own conceptions of justice. Thus, the aim is to identify those decisions that are most at odds with a court's collective judgment so that the court can correct those decisions through an appellate process.[66]

More specifically, the decision-predictive algorithm can be viewed as mimicking a hypothetical but normatively appealing world where, in each case, each judge independently casts multiple independent votes under a variety of conditions (e.g., after her football team won, after her football team lost, in the morning, in the afternoon etc.) and case outcomes are determined by the aggregation of voting results. Such a world is normatively appealing for two reasons. First, it allows us to remain agnostic with respect to the value of different judges' decisions. We avoid potentially contentious debates and instead rely on the appeal of democratic principles. Second, Condorcet's Jury Theorem, the classic theorem of political science and antecedent to the "wisdom of the crowds,"[67] provides normative grounding for this approach: as long as decision-makers are, on average, making good decisions, then a world with more independent votes will generate better decisions.[68]

Of course, in reality, such a hypothetical world is unobtainable. Requiring all judges to participate multiple time in every case would be prohibitively expensive and even absurd. But we can statistically simulate such a

[66] *See* Levy, *supra* note 45, at 424–25.

[67] The "wisdom of crowds" refers to the idea that the aggregated predictions from a large group of people will often be more accurate than most individual judgments. James Surowiecki, The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations, at xiv (2004). The classic example comes from Galton, who attended a regional fair at which 800 people each guessed the weight of an ox. The average of the guesses was just one pound away from the ox's true weight. Francis Galton, *Letters to the Editor: The Ballot-Box*, 75 Nature 509, 509 (1907).

[68] David Austen-Smith & Jeffrey S. Banks, *Information Aggregation, Rationality, and the Condorcet Jury Theorem*, 90 Am. Pol. Sci. Rev. 34, 34 (1996). The central idea of Condorcet's Jury Theorem is that a group operating under majority rule is more likely to make an accurate decision than any random member of the group deciding alone. *Id.* In the classic form, Condorcet's Jury Theorem holds true so long as (1) each individual's probability of making the right decision is greater than fifty percent; and (2) the group members' votes are cast independently of one another. *Id.* The requirement that all individuals must have a probability of making the right decision greater than 0.5 can be relaxed considerably: one only need to assume that the average of the individuals' probabilities is greater than 0.5 for the central insight of Condorcet's Jury Theorem to apply. *Id.* A key feature of the theorem is that the probability that a group will make an accurate decision increases as the size of the group increases. *Id.* Intuitively, if votes tend to be correct, then more votes are better because any one vote might be randomly mistaken.

world, and we can use the results of that simulation to identify decisions that deviate from that normative ideal.

The core idea of the decision predictive algorithm is to remove the influence of factors that are randomly, or as-if randomly assigned by adjudication systems that should not have bearing on the case outcome. Consider a simple world where there are three types of cases: X, Y, and Z. Litigants in type X cases receive a favorable ruling ninety percent of the time. The ten percent unfavorable rulings occur when a type X case happens to be assigned, for example, to a particularly harsh judge, to a judge in a particularly harsh mood, or to a judge who feels pressure to deliver an unfavorable ruling after a succession of favorable rulings in his previous five cases. In contrast, litigants in type Y cases receive a favorable ruling only ten percent of the time, by virtue of assignment to a particularly lenient judge, to a judge in a particularly lenient mood, or to a judge who feels pressure to deliver a favorable ruling after a succession of unfavorable rulings in her previous five cases. Litigants in type Z cases receive favorable rulings fifty percent of the time—judicial assessments are more contested in type Z cases. Our basic contention is two-fold. First, a court should prioritize reviewing unfavorable rulings in X cases and favorable rulings in Y cases if its goal is error-correction.[69] Second, machine-learning algorithms are a powerful tool for helping courts identify X, Y, and Z cases, as well as the nearly infinite types of cases that inhabit real-world adjudication systems.[70]

B. The Simulation Technique

The technique we propose for constructing the decision predictive algorithm to identify cases for a review is a machine-learning extension of "judgmental bootstrapping." Though the term was coined by Robyn Dawes in 1971, the concept stems from the early 1900s and was developed in a variety of fields during the 1960s.[71] The core idea is that we can build "a model of an expert by regressing his forecasts against the information that he used in order to infer the rules that the expert is using."[72] Scholars in disciplines ranging from psychology, education, marketing, and finance have applied judgmental bootstrapping to contexts ranging from school admissions

[69] *See* Copus, *supra* note 46, at 629. If a court's goal is law development, Type Z cases may be the law proper target of the court's attention. *Id.* at 633.

[70] H.D. Hughes, *An Interesting Seed Corn Experiment*, 17 Iowa Agriculturist 424, 424–25 (1917).

[71] J. Scott Armstrong, *Judgmental Bootstrapping: Inferring Experts' Rules for Forecasting*, *in* J.S. Armstrong, Principles of Forecasting 171, 173 (J. Scott Armstrong ed., 2001).

[72] Hannah Laqueur & Ryan Copus, *Synthetic Crowdsourcing: A Machine-Learning Approach to Inconsistency in Adjudication* (Dec. 6, 2017) (manuscript at 11), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2694326.

decisions,[73] predicting loan defaults,[74] criminal sentencing appeals decisions,[75] forecasting the number of advertising pages a magazine will sell,[76] and making draft picks in sports leagues.[77]

The use of multivariate regression to model an expert's reasoning process can be effective. Indeed, studies suggest that judgmental bootstrapping models outperform the experts' actual judgments in a wide variety of contexts.[78] However where decision-making requires nuanced judgments, its shortcomings can be stark.[79] Most importantly, we often have no ability to measure much of the information that judges use in a complex decision task.[80] Without that information, judgmental bootstrapping will often be unable to accurately model decision-making.[81]

We instead suggest using machine learning methods. A machine-learning approach aims not to model the expert's decision rules, but instead to produce results that make it only seem as if it has discovered the expert's decision rules.[82] The aim is to merely predict how a collective court would decide a case. Machine learning algorithms can search over rich combina-

[73] Robyn M. Dawes, *A Case Study of Graduate Admissions: Application of Three Principles of Human Decision Making*, 26 Am. Psychol. 180, 183–85 (1971).

[74] Rashad A. Abdel-Khalik & Kamal M. El-Sheshai, *Information Choice and Utilization in an Experiment on Default Prediction*, 18 J. Acct. Res. 325, 342 (1980).

[75] Duncan I. Simester & Roderick J. Brodie, *Forecasting Criminal Sentencing Decisions*, 9 Int'l J. Forecasting 49, 60 (1993).

[76] Kesten C. Green & J. Scott Armstrong, *Demand Forecasting: Evidence-Based Methods*, (Oct. 2012) (manuscript at 6–7), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3063308.

[77] Armstrong, *supra* note 71, at 172.

[78] *Id.* at 178.

[79] Laqueur & Copus, *supra* note 72, at 11; Simester & Brodie, *supra* note 75, at 49.

[80] Fred Collopy et al., *Expert Systems for Forecasting*, *in* Armstrong, *supra* note 71, at 17.

[81] Armstrong, *supra* note 71, at 174.

[82] This distinction between the linear model of parametric bootstrapping and the machine learning predictive approach are at the heart of recent shifts in computer scientific thinking with respect to Artificial Intelligence. Take, for example, the development of Google Translate. As one organizers of the Watson project explain: "the traditional symbolic AI approach to translating documents from one language to another proved ineffective. There are simply too many exceptions to the 'rules' governing how humans use languages for it to be practical to try to capture them all. . . . What big data and statistical machine learning techniques have shown us is that given enough data many of these problems can be solved to a large degree, absent deep understanding, by looking for patterns in the data." *Big Data, Statistical Machine Learning Tech-*

tions of variables, finding the combination that makes predictions most accurate.[83]

Importantly, we dispense with any restriction that the model includes only variables actually used by the decision-makers, and we instead exclude only the noise variables—those variables that we have good reason to believe are statistically unrelated to the actual merits of a case. These variables may include the judge to which one happens to be assigned, the results of the immediately preceding cases, the time of day, whether the judge's football team won the night before, the weather, and the judge's mood. By excluding these variables, predictions are averaged over the arbitrary circumstances that often influence case outcomes.

Building a model with machine learning rather than traditional linear regression methods not only allows for better predictions, but it also has the benefit of separating the predictive task from potentially controversial normative choices. The traditional regression approach puts modeling choices in the hands of the analyst. Different statistical models may generate different outcomes and thus leaves room for manipulation, whether or not intentional. A machine learning approach lets the data determine which model or combination of models generates the best predictions.[84]

C. Implementation: Machine-Guided Triage

Once a prediction algorithm is constructed, it can then be used to generate case-specific predictions of a favorable ruling, with predictions ranging between 0 and 1. By simply taking the absolute difference between the actual decision (e.g. a grant of parole coded as 1 or denial of parole coded as 0) and the prediction (e.g. a predicted grant rate of 0.1 parole), courts could rank decisions for prioritized review, with decisions having the largest difference at the top of the list. We refer to that difference as a decision's estimated "degree of error,"[85] which represents the estimated percentage of judges that would disagree with a decision. For example, imagine there is an asylum case with a 0.95 predicted probability of asylum being granted. If such a case were indeed granted, it would have an estimated degree of error of 0.05 (1-0.95). Such a low score is an indication that there is little reason to prioritize appellate review of that decision—most judges, most of the time, would agree that the asylum should be granted. But if asylum were instead denied? The decision would have a high estimated degree of error of 0.95, an indica-

niques, and Machine Translation, LA. TECH WATSON (2014), http://watson.latech.edu/book/intelligence/intelligenceOverview5b1.html.

[83] *See* Copus, *supra* note 46, at 637–38.

[84] We recommend going even further to assure model impartiality and transparency. Adjudication systems should hold public competitions and evaluate algorithms solely on their predictive capacity. For discussion of this point, *see id.* at 661.

[85] *Id.* at 629.

tion that something has gone wrong and that the decision should be prioritized for review.

This algorithmic approach could be used as a standalone tool for appellate screening or combined with traditional screening tools—filing fees and preliminary assessments—that are already in place. For example, courts could increase filing fees as the degree of error decreases so as to discourage meritless appeals while still providing confident litigants with a way to override an inaccurate algorithmic assessment. Courts could also intensify the resources and attention spent on their preliminary assessment as the degree of error increases. Courts could thus make sure to catch those cases with a high degree of error, while still relying on human intervention to remedy obvious errors that are incorrectly given low error scores by the algorithm.

Note that, in some adjudication systems, the set of judges that make initial decisions and the set of the judges that review those initial decisions are not the same judges.[86] Although the ultimate goal is to identify cases that the reviewing judges would reverse, the decision predictive algorithm should nonetheless be constructed using the collective judgment of the lower-level judges. The reason is two-fold. First, the larger sample size of the lower courts will allow for more accurate predictions.[87] Second, and most importantly, modeling the lower court allows for a match between the sample of cases used to build the model and the set of cases to which the model is applied (i.e., the lower-court cases). In contrast, if the model were built on appellate cases, there would be a mismatch between the data use to build the model and the population of lower-court cases that the model is applied to. There would be little reason to expect that a model built on appellate cases could be usefully applied to the set of lower-court cases. Thus, even where there are moderate departures between judicial views in the two levels, these benefits still counsel modeling the decisions of the lower court.[88]

D. The Virtues of Decision Predictive Algorithms as a Screening Tool

Using predictive models of judicial decisions to prioritize cases for review effectively overcomes the standard objections to the use of algorithms

[86] *Introduction to the Federal Court System*, *supra* note 2.

[87] Supervised machine learning algorithms are "data hungry." *See, e.g.*, Ignacio Olmeda & Pauline J. Sheldon, *Data Mining Techniques and Applications for Tourism Internet Marketing*, 11 J. Travel & Tourism Marketing 1, 16 (2008). That is, they work most effectively by sifting through large numbers of variables, looking for combinations that reliably predict outcomes. They therefore demand more data than parametric techniques and generally, the more data available the better the algorithm will perform. Alon Halevy et al., *The Unreasonable Effectiveness of Data*, 24 IEEE Intelligent Sys. 8, 8 (2009).

[88] Where the lower and upper court views depart significantly, there is less need for any type of screening—insofar as the lower court is systematically at odds with the higher court, any appeal is more likely to result in a reversal.

in the legal system. First, it significantly mitigates the problems of technical bias.[89] Recall that models based on events external to an adjudication system, like those that predict whether an individual will offend, are plagued by problems of measurement. If the crimes of some groups are more likely to be detected than other groups, their algorithmically-assessed risk will be artificially inflated. There is no such mismeasurement problem when the target of the prediction is a judicial decision—administrative records should make it trivial to accurately collect data on the outcomes of cases. Moreover, because we can observe an outcome for each case in an adjudication system, a model of decisions can bypass the selective labels problem.[90]

Second, employing algorithms to screen for likely errors rather than to guide merit decisions mitigates due process concerns.[91] As a matter of logic, the appellate screening process is relatively light on due process: a comprehensive, individualized assessment of each case's merits would amount to a full appellate hearing for each case. But a fundamental function of the appellate process is to preserve its superior resources for a more finely screened set of cases. Whether that screening occurs through the imposition of costs that encourages litigants to self-screen, through preliminary assessments that "take a peek" at case merits, or an algorithm that statistically summarizes case probabilities of reversal makes little difference in terms of due process.

Third, when algorithms are used as an appellate screening mechanism, courts can freely share them with the public. While litigants might be inspired to game algorithms if it provides them with an automated decision or recommendation, there is significantly less benefit in attempting to game algorithms if the ultimate benefit is simply access to secondary review. A low merit case would be even less likely to succeed at the appellate level than at the initial stage, so the benefit of review would be insubstantial. With minimal concerns about litigant gaming, adjudication systems would have little justification for keeping algorithms secret.[92]

[89] Lakkaraju et al., *supra* note 29 (technical bias occurs when the dataset used for the algorithm is not congruient with the population being assessed). When applied to prioritize cases, the dataset is the outcome of cases.

[90] For a discussion of the selective labels problem, see *supra* Section II.B.

[91] Freeman, *supra* note 18, at 140 (explaining that due process concerns typically arise regarding the use of algorithms in pre-sentencing calculations). But, when employed to screen for errors, the time at which due process concerns regarding algorithms has already passed.

[92] Trade secrets are currently the most common justification for secrecy. But government now has the ability to access high-quality algorithms through public competitions. For example, organizations like Netflix, Homeland Security, and Microsoft have hosted competitions at Kaggle.com. John Mannes, *The Kaggle Data Science Community is Competing to Imporove Airport Security with AI*, TechCrunch (June 22, 2017), https://techcrunch.com/2017/06/22/the-kaggle-data-science-community-is-competing-to-improve-airport-security-with-ai/.

Fourth, models of decision-making can be constructed for any adjudication system. As noted above, most adjudication systems do not have access to an external outcome measure like criminal offending upon which to build an algorithm that might guide decisions. But every adjudication system should have a dataset of its own decisions, allowing for easy expansion of algorithmic aid to almost any adjudication system.[93]

Finally, using algorithms to target cases for secondary review is largely immune to the substantive allegation that algorithms can embed and cement biases even when human decision-makers may evolve and change.[94] Because all merit decisions are ultimately in the hands of human decision-makers, the algorithm can easily update with evolving human judgment: as long as algorithms are trained to be most predictive of contemporary decisions, models will be only minimally tied to historical norms that a society may ultimately conclude are biased and outdated. As judicial norms change, the data on which algorithms are trained—the initial decisions of judges—can change without being hampered by the guidance of an algorithm.

IV. CONCLUSION

Adjudication systems, notably in the area of criminal justice, are increasingly turning to data-driven algorithms in the hopes of improving decision-making. However, there are also growing concerns that algorithm-guided decision making can embed racial biases, deprive parties of due process, and impair transparency. Rather than use algorithms to recommend or even automate decisions, we have proposed using algorithms to reboot an older and well-established system for improving the quality of legal decisions: appellate review. Implementing algorithms to target decisions for ap-

There thus seems little reason for relying on proprietors who claim trade secret protections.

[93] Some adjudication systems will face complications. We have two situations in mind. First, the decision that an appellate court faces may be different than the decision faced by the lower court due to deferential standards of review. The challenge could likely be overcome by maintaining separate thresholds for cases with different standards of review. For example, perhaps decisions that would be reviewed de novo should be prioritized for review if they have a degree of error greater than 0.5, while decisions facing a clearly erroneous style should only be prioritized for review if they have a degree of error greater than 0.8. A second issue is presented when courts do not record the results of settlement. When lacking measurements of the outcome for certain cases, it makes it more difficult to generate models of outcomes that are used to generate the degree of error.

[94] Lakkaraju et al., *supra* note 29 (technical bias occurs when the dataset used for the algorithm is not congruent with the population being assessed). By training the algorithms to be most predictive of more recent cases, the algorithm will naturally update as the judicial preferences evolve.

pellate review can still leverage the core benefits of predictive technology while avoiding the most potent objections.