Machine learning and the rule of law

Daniel L. Chen

Machine Learning and the Rule of Law

Daniel L. Chen*

*To whom correspondence should be addressed.

Predictive judicial analytics holds the promise of increasing the fairness of law. Much empirical work observes inconsistencies in judicial behavior. By predicting judicial decisions—with more or less accuracy depending on judicial attributes or case characteristics—machine learning offers an approach to detecting when judges most likely to allow extralegal biases to influence their decision making. In particular, low predictive accuracy may identify cases of judicial "indifference," where case characteristics (interacting with judicial attributes) do no strongly dispose a judge in favor of one or another outcome. In such cases, biases may hold greater sway, implicating the fairness of the legal system.

Introduction

There is ample social scientific evidence documenting arbitrariness, unfairness, and discrimination in the U.S. legal system. To give just a flavor

*Daniel L. Chen, daniel.chen@iast.fr, Toulouse School of Economics, Institute for Advanced Study in Toulouse, University of Toulouse Capitole, Toulouse, France; Directeur de Recherche, Centre National de la Recherche Scientifique. Work on this project was conducted while Chen received financial support from the European Research Council (Grant No. 614708), Swiss National Science Foundation (Grant Nos. 100018-152678 and 106014-150820), and Agence Nationale de la Recherche.

of the relevant research:

- U.S. federal appeals court judges become more politicized before elections and more unified during war (Berdejo and Chen 2016; Chen 2016b).
- Refugee asylum judges are 2 percentage points more likely to deny asylum to refugees if their previous decision granted asylum (Chen, Moskowitz, and Shue 2016).
- Politics and race also appear to influence judicial outcomes (Schanzenbach 2005; Bushway and Piehl 2001; Mustard 2001; Steffensmeier and Demuth 2000; Albonetti 1997; Thomson and Zingraff 1981; Abrams, Bertrand, and Mullainathan 2012; Boyd, Epstein, and Martin 2010; Shayo and Zussman 2011) as does masculinity (Chen, Halberstam, and Yu 2016b, 2016a), birthdays (Chen and Philippe 2017), football game outcomes (Chen 2017; Eren and Mocan 2016), time of day (Chen and Eagel 2016; Danziger, Levav, and Avnaim-Pesso 2011), weather (Barry et al. 2016), name (Chen 2016a), and shared biographies (Chen et al. 2016) or dialects (Chen and Yu 2016).
- There are also various papers showing clear judicial biases in laboratory environments, such as the influence of anchoring, framing, hindsight bias, representative heuristics, egocentric bias, snap judgments, and inattention (Guthrie, Rachlinski, and Wistrich 2000, 2007; Rachlinski et al. 2009; Rachlinski, Wistrich, and Guthrie 2013; Simon 2012).

Thus, the primary question is not whether these problematic features of the legal system exist. Rather, the dilemma facing policymakers is what, if anything, can be done. This comment will argue that predictive judicial analytics in the form of applied statistical/machine learning (from causal inference to deep learning) holds at least some promise on this front.

Prior empirical work has focused on evaluating judges to observe the influences on their behavior, helping to diagnose the problem of bias but offering little in terms of remedy. The advent of machine learning tools and their integration with legal data offers a mechanism to detect in real time, and thereby remedy judicial behavior that undermines the rule of law. This commentary presents a conceptual framework for understanding a large set of behavioral findings on judicial decision-making and then taking steps to ensure more fair treatment of legal subjects by the legal system.

The theoretical basis for the following argument is the observation that behavioral biases are most likely to manifest in situations where judges are closer to indifference between options. Such contexts are also those where there are likely to be the highest levels of disparities in inter-judge accuracy of algorithms predicting judicial decisions—essentially conditions where judges are unmoved by legally relevant circumstances. If algorithms can identify the contexts that are likely to give rise to bias, they can also reduce those biases through behavioral nudges and other mechanisms, such as through judicial education. The following discussion fleshes out these claims.

The Problem of Indifference

Imagine a legal outcome of interest, such as asylum designations by immigration judges. Let's denote that outcome Y. Imagine further that there is some set of covariates (or "features" in the language of machine learning) X such that these features can be used to generate predictions of Y via some function $Y = f(X) + \epsilon$ where ϵ denotes some small "error" or variation. The covariates X are *legally relevant* in as much as prevailing legal norms require or least permit their use by legal decisionmakers for the relevant decision. In the case of an asylum adjudication, the political circumstances of an applicant's home country would be legally relevant.

There might also be a set of covariates W that are legally irrelevant, and that *should not* predict a legal outcome: $y \perp W, var(\varepsilon) \perp W$. The

set W might include litigant characteristics that decisionmakers are not permitted to take into account—such as race—or they may include irrelevant features in the environment, such as the weather. Since judges are randomly assigned, judicial characteristics fall into W, because the judge that a litigant randomly draws is not legally relevant to the outcome of a decision. Of course, as mentioned briefly above, there is a substantial literature showing that features in W in fact are predictive of legal outcomes in a variety of settings (Berdejo and Chen 2016; Chen 2016b; Chen, Moskowitz, and Shue 2016; Schanzenbach 2005; Bushway and Piehl 2001; Mustard 2001; Steffensmeier and Demuth 2000; Albonetti 1997; Thomson and Zingraff 1981; Abrams, Bertrand, and Mullainathan 2012; Boyd, Epstein, and Martin 2010; Shayo and Zussman 2011; Chen and Philippe 2017; Eren and Mocan 2016; Chen 2017; Chen and Eagel 2016; Danziger, Levav, and Avnaim-Pesso 2011; Barry et al. 2016; Chen et al. 2016; Chen and Yu 2016).

The preferences of decisionmakers (e.g., judges) over X may also affect the influence of W over outcomes. A judge could be said to have *strong* preferences over X when it is costly to depart from the legally optimal outcome, defined as the outcome that would be generated through consideration of X alone. Judges might have such strong preferences based on ideology, or personal psychology, or some set of institutional characteristics. But a judge may also have *weak* preferences over X, meaning that there was a relatively low cost in departing from the legally optimal outcome. In such cases of *legal indifference*, the factors within W can be expected to have greater influence. Stated another way, when the predictive power of X wanes, the potential scope of influence for W waxes.

A Role for Machine Learning

Chen et al. (2017) conceptualize the notion of early predictability. The basic idea is that machine learning could be used to automatically detect judicial indifference—i.e., instances where the judges appear to ignore the circumstances of the case when making decisions. This infor-

mation could then be used to trigger de-biasing information or other interventions to prevent decisions that would undermine the fair and non-arbitrary operation of the justice system.

How would this work? Continuing our example, let's consider asylum courts. In this important context it turns out that using only the information that is available at the time that a case opens, judges (those with the highest and lowest grant rates) are much more predictable than others (Chen et al. 2017). These judges seem to have strong prior preferences concerning outcomes and the legally irrelevant fact that an applicant is assigned to a low- or high-grant rate judge largely determines outcomes.

There is, however, a category of less predicable judges and these judges tend to have middling grant rates. Given their unpredictability, one possibility is that they lack strong preferences, and are therefore guided by random factors when making a decision—essentially flipping a coin. Another possibility is that they are more sensitive to the circumstances of the cases. There is some evidence pointing to this second alternative: the less predictable judges also tend to have substantially more hearing sessions than the judges who rarely grant asylum.[1]

At this level of granularity—identifying judges whose behavior is predictable at relatively early procedural stages—some interventions might be possible. For example, training programs could be targeted toward these judges, either with the goal of de-biasing or to help them learn how to use the hearing process to better advantage. Simply alerting judges to the fact that their behavior is highly predictable in ways that may indicate unfairness may be sufficient to change their behavior.

Higher levels of granularity in the analysis may free up even more targeted interventions. It may be possible, for example, to not only identify early deciding judges, but also to examine how case characteristics inter-

1. Interestingly, judges who grant at a high rate also hold relatively more hearing sessions, perhaps to collect more information to justify their likely more controversial decision.

act with this judicial attribute to learn the types of case-judge pairs that are most predictable at early stages. When such pairs are found, judges can be given a 'red flag' that they should be particularly attuned to subsequent information, essentially as a counter-weight to confirmation bias or other non-legal sources of influence.

Just as machine learning can be used to identify judges who tend to be unmoved by legally relevant factors, these techniques can also be used detect instances where judicial decisions can be predicted by legally irrelevant factors. There is a substantial social science literature establishing this possibility, and in the asylum example, Chen and Eagel (2016) finds these influences are highly prevalent. They include: whether a hearing was before lunch or towards the end of the day; the size of the applicant's family; the weather; the number of recent grants by the court; whether genocide has been in the news; and the date of the decision. While the literature typically studies one behavioral feature at a time, Chen and Eagel (2016) demonstrates the possibility for machine learning to automate the detection of inconsistencies between judges due to legally irrelevant factors.

The asylum example is just one of many where machine learning techniques can be use to detect bias. Amaranto et al. (2017) uses a very large data set concerning prosecutorial decisions in New Orleans over a twelve-year period (430,000 charges and 145,000 defendants) to test for racial bias and the effectiveness of prosecutors in pursuing the riskiest defendants. The authors construct a predictive model of recidivism and then test if prosecutors who are relatively strict (i.e., drop relatively few cases) screen based on risk of recidivism. In fact, more stringent prosecutors do not appear to target riskier defendants, and this phenomenon has differential effects across races, with less risky African-American defendants actually receiving relatively harsh treatment.

More generally, machine learning techniques can be used on data of any sort, and in the context of a legal decision, a wide range of data, from the weather conditions to judge characteristics, have proven informative.

Given the textual nature of the law, and the importance of argumentation and reason-giving to legal decision making, there is a substantial amount of textual data that can be used to examine how legally relevant and legally irrelevant factors affect legal outcomes. For example, Ash and Chen (2017) use judges' writings to predict the average harshness and racial and sex disparities in sentencing decisions. That work finds that the information contained in written opinions can improve significantly on naive prediction of punitiveness and disparity.

Again, this information could be used to aid decision makers in ways that reduce bias in the system. Informing judges about the predictions made by a model decision maker could help reduce judge-level variation and arbitrariness. Potential biases that have been identified in prior decisions or writing could be brought to a judge's attention, where they could be subjected to higher order cognitive scrutiny. Such efforts would build on the already significant push to integrate risk-assessment into the criminal justice process to help inform judges of the objective risks posed by defendants.

Judicial Education

An additional pathway for machine learning to improve the quality of legal decision making is by informing, and to some extent comprising, efforts at judicial education. The first goal would be to expose judges to findings concerning the effects of legally relevant and legally irrelevant factors on decisions, with the goal of *general* rather than *specific* debiasing. For example, Pope, Price, and Wolfers (2013) found that awareness of racial bias among NBA referees subsequently reduced that bias. The second goal would be to educate legal decision makers in the tools of data analysis, so that they can become better consumers of this information when it is present during legal proceedings, and to more generally provide a set of thinking tools for understanding inference, prediction, and the conscious and unconscious factors that may influence their decision making.

Efforts at judicial education have had considerable success in the past. By 1990, 40% of federal judges had attended an economics-training program. This law and economics program was founded in 1976 as a two-week training course with lectures by Nobel Prize economists Milton Friedman, Paul Samuelson, and other luminaries. Ash, Chen, and Naidu (2017) tests for effects from this training, finding dramatic results. Economics language used in academic articles become rapidly prevalent in judicial opinions. Economics training affects both the trained judges and their peers as economic language travels from one judge to another and across legal areas. Perhaps most tangibly, economics training changed how judges perceived the consequence of their decisions. Judges in economic regulation cases shifted their votes in an anti-regulatory direction by 10%. In the district courts, when judges were given discretion in sentencing, economics trained judges immediately rendered 20% longer sentences relative to the non-economics counterparts.

Part of what made the economics training program successful is likely because theory provided structure for judges to understand the patterns they saw. The question for theorists and researchers now is whether machine learning, text-as-data analysis, and other similar developments allow for a further step. If judges are shown the behavioral findings, will they become less prone to behavioral biases? If judges are taught theoretical structure that drive the behavioral bias, will they become better judges? Could a new generation of theory and evidence from behavioral and social sciences provide better justice and increase cooperation, trust, recognition and respect?

References

Abrams, David, Marianne Bertrand, and Sendhil Mullainathan. 2012. "Do Judges Vary in Their Treatment of Race?" *Journal of Legal Studies* 41 (2): 347–383.

Albonetti, Celesta A. 1997. "Sentencing under the federal sentencing guidelines: Effects of defendant characteristics, guilty pleas, and departures on sentence outcomes for drug offenses, 1991-1992." *Law and Society Review:* 789–822.

Amaranto, Daniel, Elliott Ash, Daniel L. Chen, Lisa Ren, and Caroline Roper. 2017. "Algorithms as Prosecutors: Lowering Rearrest Rates Without Disparate Impacts and Identifying Defendant Characteristics 'Noisy' to Human Decision-Makers."

Ash, Elliott, and Daniel L. Chen. 2017. "Predicting Punitiveness from Judicial Corpora."

Ash, Elliott, Daniel L. Chen, and Suresh Naidu. 2017. "Ideas have Consequences: The Impact of Law and Economics on American Justice." April.

Barry, Nora, Laura Buchanan, Evelina Bakhturina, and Daniel L. Chen. 2016. "Events Unrelated to Crime Predict Criminal Sentence Length."

Berdejo, Carlos, and Daniel L. Chen. 2016. "Electoral Cycles Among U.S. Courts of Appeals Judges." *The Journal of Law and Economics* 60 (3): 479–496.

Boyd, Christina L., Lee Epstein, and Andrew D. Martin. 2010. "Untangling the causal effects of sex on judging." *American Journal of Political Science* 54 (2): 389–411. doi:10.1111/j.1540-5907.2010.00437.x.

Bushway, Shawn D, and Anne Morrison Piehl. 2001. "Judging judicial discretion: Legal factors and racial discrimination in sentencing." *Law and Society Review:* 733–764.

Chen, Daniel L. 2016a. "Implicit Egoism in Courtrooms: First Initial Name Effects with Randomly Assigned Defendants."

———. 2016b. "Priming Ideology: Why Presidential Elections Affect U.S. Judges."

———. 2017. "Mood and the malleability of moral reasoning."

Chen, Daniel L., X. Cui, L. Shang, and J. Zhang. 2016. "What Matters: Agreement Between U.S. Courts of Appeals Judges."

Chen, Daniel L., Matt Dunn, Levent Sagun, and Hale Sirin. 2017. "Early Predictability of Asylum Court Decisions." *Artificial Intelligence and the Law* (March).

Chen, Daniel L., and Jess Eagel. 2016. "Can Machine Learning Help Predict the Outcome of Asylum Adjudications?" *Artificial Intelligence and the Law* (March).

Chen, Daniel L., Yosh Halberstam, and Alan C L Yu. 2016a. "Covering: Mutable Characteristics and Perceptions of (Masculine) Voice in the U.S. Supreme Court."

———. 2016b. "Perceived Masculinity Predicts United States Supreme Court Outcomes." *PLOS-ONE*, *11(10)*, *e0164324*.

Chen, Daniel L., Tobias J Moskowitz, and Kelly Shue. 2016. "Decision-Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires." *The Quarterly Journal of Economics* 131 (3): 1181–1241.

Chen, Daniel L., and Arnaud Philippe. 2017. "Clash of Norms: Judicial Leniency on Defendant Birthdays."

Chen, Daniel L., and Alan Yu. 2016. "Mimicry: Phonetic Accommodation Predicts U.S. Supreme Court Votes."

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. 2011. "Reply to Weinshall-Margel and Shapard: Extraneous factors in judicial decisions persist." *Proceedings of the National Academy of Sciences of the United States of America* 108 (42): E834.

Eren, Ozkan, and Naci Mocan. 2016. "Emotional Judges and Unlucky Juveniles." *American Economic Journal: Applied Economics* 10 (3): 171–205.

Guthrie, Chris, Jeffrey J Rachlinski, and Andrew J Wistrich. 2000. "Inside the Judicial Mind." *Cornell Law Review* 86 (4): 777–830.

———. 2007. "Blinking on the Bench: How Judges Decide Cases." *Cornell Law Review* 93 (1): 1–44.

Mustard, David B. 2001. "Racial, ethnic, and gender disparities in sentencing: Evidence from the US federal courts." *Journal of Law and Economics* 44 (1): 285–314.

Pope, Devin G, Joseph Price, and Justin Wolfers. 2013. *Awareness reduces racial bias*. Technical report. National Bureau of Economic Research.

Rachlinski, Jeffrey J, Sheri Lynn Johnson, Andrew J Wistrich, and Chris Guthrie. 2009. "Does Unconscious Racial Bias Affect Trial Judges?" *Notre Dame Law Review* 84:1195–1246.

Rachlinski, Jeffrey J, Andrew J Wistrich, and Chris Guthrie. 2013. "Altering Attention in Adjudication." *UCLA Law Review* 60:1586–1618.

Schanzenbach, Max. 2005. "Racial and sex disparities in prison sentences: the effect of district-level judicial demographics." *The Journal of Legal Studies* 34 (1): 57–92.

Shayo, Moses, and Asaf Zussman. 2011. "Judicial Ingroup Bias in the Shadow of Terrorism." *The Quarterly Journal of Economics* 126 (3): 1447–1484.

Simon, Dan. 2012. *In Doubt: The Psychology of the Criminal Justice Process.* Cambridge, MA: Harvard University Press.

Steffensmeier, Darrell, and Stephen Demuth. 2000. "Ethnicity and sentencing outcomes in US federal courts: Who is punished more harshly?" *American sociological review:* 705–729.

Thomson, Randall J, and Matthew T Zingraff. 1981. "Detecting sentencing disparity: Some problems and evidence." *American Journal of Sociology:* 869–880.