Judge, Jury, and EXEcute file: The Brave New World of Legal Automation

By Elliott Ash
June 2018

While (the option of) replacing a judge with a robot is still many years away, the technological trends that will lead to this eventuality have been in motion for decades. As early as the 1970s, computerized legal research services began to transform legal practice. Automated search through legal databases eventually gave way to automated search through digital document evidence. And most recently, legal automation is taking over contract review and drafting. In the courts, we have even begun to see the use of algorithms in decisions about whether to grant bail or parole.

This report discusses the prospects for automating decisions in the legal system. I will discuss active research on decision prediction models for judges and prosecutors and how these algorithms might be used to detect and reduce bias in legal decision-making. I will also discuss the substantial risks for these algorithms to replicate existing biases in the system or create new ones. Along the way, I will discuss the role that incentives theory and econometrics can play in understanding and mitigating these risks.

Predicting Legal Decisions

Let us start by discussing a simplified description of the legal decision process. The robot judge faces a decision, which could in principle include anything that judges, juries, or other juridical entities decide on a routine basis. It could be a decision whether to dismiss a case, whether to grant money bail, or whether to declare a mistrial. It could be a case disposition: guilty or innocent, or the length of a prison sentence. In all these cases, the decision can be reduced to a selection among some set of options, or a numerical value.

The robot judge uses evidence, which could include anything that is machine-readable. For starters, this would include digital text, such as digitized police reports, deposition and trial transcripts, affidavits, appellate briefs, and the like. The literature on computational linguistics is now quite advanced and can produce rich statistical representations of written language (Jurafsky and Martin, 2014). The technology for

audio and visual processing, while more recent, has also made astounding strides. Therefore the evidence used by the robot judge would not be limited to texts. It could also include the audio, image, or video files from police cameras, depositions, and trials.

With some measure of evidence and some measure of the decision, it is straightforward in principle to automate legal decision-making. First, one collects data on some or all of the previous cases, which consist of evidence-decision pairs. These pairs then become a data set for the machine learning model.

The model would then work to produce an approximation of the legal rules used by judges. If the approximation is close to the truth, then many cases will be predicted correctly. If it is far from the truth, then many cases will be predicted incorrectly. Machine learning models are designed to gradually update the estimate to move gradually closer to the true legal rule. Eventually, the model will get as close as possible given the available evidence data.

There are many machine learning models being used by researchers and industry, and they are available off the shelf as open source software.[1] What these models share is a set of learning rules that extract the predictive information from a training set without overfitting the data. Features that are "useful" for the prediction task (minimizing errors) are automatically identified and up-weighted by the model, features that are poor predictors are similarly identified and ignored.

In the case of the law, these models would take a set of case characteristics (evidence) and tell us how a judge would probably decide. This is a small but active research area. These papers include Ash et al. (2018) (predicting bankruptcy court decisions with 67 percent accuracy and minimal case information), Katz et al. (2017) (predicting U.S. Supreme Court decisions with 70 percent accuracy), Chen and Eagel (2017) (predicting asylum courts with 82 percent accuracy), and Amaranto et al. (2017) (predicting prosecutor charge decisions with 88 percent accuracy).

All of these predictions were made on publicly available data that the court collected for administrative purposes. A legal system that intentionally collects information for automation purposes would produce much richer, reliable, and predictive data sets. Further, this literature is somewhat young, and I am confident that these predictions will continue to improve over time. If the model predicts with good accuracy on the existing set of cases, then we might be confident enough to use the predictions for new cases and apply them as law.

---

[1] See, e.g., the Scikit-learn package in Python (e.g. Pedregosa et al., 2011). For classification tasks, a popular model is regularised logit, for regression tasks, one might use elastic net. Other effective models are random forests, boosted trees, and neural networks

Promises of Automated Adjudication

In the end, what we hope to have is a case prediction machine that could assist judges in their jobs. While the algorithm would learn to minimize errors, it would never be perfect. But human judges are not perfect, and the system has the major promise that it could help reduce human error.

There is a wealth of evidence for the shortcomings of human judges. First, judges vary widely in their decision tendencies when faced with the same facts. For example, Keith et al. (2013) document the large variation across asylum judges in the probability of refugees being granted asylum in the United States, holding all facts constant. Further, judges sometimes respond to extraneous factors such as hunger, weather, and sporting event outcomes (Danziger et al., 2011, Chen, 2014).

This means that there is no "true" set of legal rules specifying the operation of the legal system. Every judge has their own set of rules and procedures. And even within the same judge, those rules change over time, sometimes in response to extraneous factors that most feel should not be considered. There is already some arbitrariness in the system, where given the exact same evidence, there is some probability for both a "yes" and "no" decision. Even if these extraneous factors did not prejudice any particular group of people, the resulting randomness could be a problem to the extent that one values consistent application of the laws.

The machine learning approach can address this problem. If the algorithm does not use the assigned judge or the weather on trial day as variables, then they will not affect the decision. That is, the model would always produce the same decision for the same set of evidence. This reform might then result in fewer judicial errors. Any individual judge biases (that is, biases of individual judges that are netted out when averaging over all cases) would be corrected.

Practically speaking, I envision a short-term implementation of this tool as a sort of robot clerk. It is an app that takes in evidence data, runs the numbers, and then produces a prediction about what previous judges would have likely decided. Human judges would then use this prediction as an input into their own decision, which could be based on a wider range of factors. In uncertain cases (near 50/50), the decision would still be made wholly by the human judge.

Many, probably most, cases in modern legal systems are clear-cut on the merits (e.g. Posner, 2008). A clerk with an hour or two of time could figure them out easily. But in many legal systems, including many state courts in the United States, the judges and clerks are overwhelmed with a massive caseload, and the clerks do not have these two hours to spare. If the algorithm can figure these cases out in a second or two, that would save a lot of time. If it improved consistency of decisions across judges in the meantime, that would be an added benefit.

Predicting Outcomes Rather than Decisions

The approach outlined so far is based on predicting legal decisions based on past decisions. An alternative use of automation in the legal system is for predicting outcomes (rather than decisions). For example, a judge might want to know whether a defendant is likely to commit other crimes. Or whether the pollution emitted by a company is likely to cause injuries. As of now, this type of information might be generated by expert testimony. But here we imagine a set of policies for more systematic prediction of relevant outcomes of judicial decisions.

The machine learning approach is the same as before. But rather than predicting a decision, one predicts some associated outcome of a decision. This outcome could be whether the defendant in a case is re-arrested for a subsequent crime, for example. The goal, again, is to minimize the number of mis-predicted outcomes. If the error rate is low enough, then the resulting prediction could be sensibly used to guide decision-making. In principle, any measurable outcome that would be useful for a judge could be predicted.

In practice, one of the main applications of this type of approach has been in criminal risk scoring. For example, Kleinberg et al. (2017) predict whether a defendant released on bail is subsequently re-arrested, and then compare the machine-predicted decisions to the decisions of human judges. Amaranto et al. (2017) undertake a similar analysis in the case of screening prosecutors. In both of these papers, it is showed that the machine help human judges bail criminality risk.

Bias and Fairness

So far we have discussed the promises of automated prediction of case decisions and outcomes. As discussed, these algorithms would provide predictions about what the average judge would do based on the evidence. Any individual judge biases would be corrected. A more serious problem with the approach outlined here is that judges are biased on average.

Most of the research on systematic bias in the judiciary is on how black defendants are treated in the U.S. criminal justice system, holding other factors constant (Fagan and Ash, 2017). Black defendants are stopped more often, given bail less often, are charged with more serious crimes for the same acts, are more likely to be convicted by juries, and given longer sentences by judges. These many biased legal decisions, taken together, also tend to result in disparities in socioeconomic outcomes (Alexander, 2012). In consequence a system with initially biased treatment could result in continued biased outcomes even when the initially biased procedures are corrected.

This matters for the robot judge because any automated decision system that is trained on biased data will also be biased. Whether currently implemented criminal risk scores,

such as COMPAS, are in practice biased is an area of active investigation and debate. A major challenge is that many currently used risk metrics are proprietary, closed-source, and developed by for-profit companies motivated to defend the validity of the scores. Skeem and Lowenkamp (2016) examine a risk metric used in federal courts and find that while blacks and whites who are otherwise identical will be treated the same, blacks tend to be rated as more risky due to longer criminal histories. Given the pre-existing biases in the criminal justice system, this is exactly how automated risk metrics can reproduce bias.

In general, there is no good way to fix these problems. If we ignore race, then blacks are still being discriminated against due to the presence of criminal history. If we ignore criminal history, then the model would become less accurate, putting more innocent people in jail and letting more guilty people go free.

Complexity, Obscurity, and Transparency

A well-known problem with legal machine learning tools is that the models are complex. They can take dozens or even hundreds of facts and characteristics into account. With these complex models, it is unclear why any particular decision is made. A downside of complexity and obscurity is that some individuals will have a better understanding of the algorithm than others. Individuals with preferential knowledge would have an advantage in preparing evidence to submit to the system.

The issue of preferential knowledge could be worse if the workings of the algorithm are not publicly known. If the algorithm is proprietary or otherwise private, then insiders might know more about it than outsiders. A closed-source algorithm (as are some current criminal risk metrics, such as COMPAS) has the potential for unchecked unfairness and abuse. An open system has the advantage that it can be examined by neutral experts.

On the other hand, an open-source algorithm might allow legal actors to game the system. Savvy attorneys (including prosecutors) could learn to produce evidence that appears innocuous but fools the algorithm into deciding one way or the other. In terms of down-stream actions, individuals might learn that some types of (legally relevant) actions are ignored by the algorithm.

A compromise transparency policy would be to make the code open source, but to make the evidence weights used by the algorithm private. In this case, the public would not know the particular action or evidence used by the algorithm, but they would be able to verify whether the parameters were learned fairly.

A final issue is that even if the code is fair, the resulting model may be faulty if it is trained on faulty data. The data sets on previous cases used for the prediction task would have to be available to experts to evaluate the system fully. For the guilty, this may not be a big deal. But to replicate the training, one would also need information on the innocent, including those falsely accused. Overall, there are major privacy issues to trade off against transparency benefits in the design of automated legal decision systems.

Reading and Explaining the Laws

There is another quite important limitation on the prediction systems outlined thus far. That is, the system would only work on legal areas already considered by many cases in the past. The machine would not work on new types of cases and, in particular, it would not account for new laws and legislation. This may be the key argument against calling this system a legal artificial intelligence, which would be a much taller order.

A legal artificial intelligence must read, understand, interpret, and apply the law. These consist of millions of pages of legislation, regulations, treaties, constitutions, judicial decisions, academic commentary, and the list goes on. Computers can process these large volumes of text easily, and technologies in computational linguistics are developing rapidly (Jurafsky and Martin, 2014). Computer programs can extract grammatical relations, can identify agents, actions, and objects, can construct basic factual relations, and can understand basic tasks. But the way computers understand language is still quite primitive compared to the way that humans understand it. And legal language is at least as complex as human language in other domains.

But let's assume for a moment that these challenges were overcome and a legal AI could explain its decision process. In this case, the computer, unlike a human, would be able to provide a true (verifiable) reason for its decision. Human judges can explain their reasoning in an opinion, but (since we cannot read their minds) it is impossible to verify independently that the reasons provided are sincere. As discussed earlier, judge decisions respond to extraneous factors such as the weather and sports games (Chen, 2014). These judges are likely unaware that these extraneous factors are affecting their decisions, and therefore would not include them in their articulated reasoning. A major advantage of a computerized judge is that the "reasons" for the decision (at least in terms of the coefficients on relevant evidence variables) would be transparent and verifiable.

Legal Vagueness and Policy Outcomes

Even if the legal AI were taught to read the laws, a thorny issue remains. Legal experts often disagree on the meaning of laws, which can be vague or indeterminate. Consider a law requiring that drivers take "reasonable precaution" on the road (many laws have this flavour). By itself, the clause is too vague to guide decisions in reckless driving cases. Human judges would decide cases based on context-dependent and case-specific details. The AI could be trained on these cases as outlined above.

But what if we are concerned that the human judges are making arbitrary or bad decisions? If we would like the robot judge to try to select among the better judgements, it would have to make value judgments. Put differently, rather than teaching the AI how current procedures work toward a decision, we might try to teach it how the decision *should* be made. For this purpose, we can imagine some social welfare criterion that depends on the decision. In the case of reckless driving, for example, this value could be the costs from traffic injuries and fatalities. In criminal law, it could be the costs of crime, which would include those associated with the current defendant as well as the deterrence impact on other potential offenders.

Given this criterion, the legal AI could design a policy to maximize it. That is, the robot judge could set up a decision process that tends to maximize the expected welfare given the evidence. The model parameters, which tell us the predicted impact of a decision on the welfare outcome, could be learned from the impacts of previous decisions as outlined earlier.

A major challenge to this approach is how to choose the welfare criterion. Welfare is subjective, and there will be disagreement about what counts as a "good" legal outcome. For now, human judges have to come to some decision based on their preferences and their beliefs about the current case and the precedential impact in future cases. But these values are private to the judge and could not easily be taught to a machine. Ideally, most controversial issues could be resolved through the democratic process. But the legal system faces so many questions, of varying importance, that it would not be feasible to have a vote on every issue by voters or legislators. A related problem is that even if a welfare criterion is specified, it may be difficult to measure. If we start rewarding measurables (e.g. quick trials), we could be trading off immeasurables (e.g. "a fair day in court").

A second challenge is how to measure the impact of decisions on the welfare criterion. Above, we outlined how to learn the impact of a bail decision on subsequent re-arrest rates. In principle this approach could also be used in more policy-based domains, even social issues such as abortion and religious freedoms (e.g. Ash and Chen, 2017). But when learning about case impacts, there is a serious issue of distinguishing correlation from causation. Legal-system outputs are high-dimensional and correlated with other external factors. As one example, the AI would not be able to determine the

causal effect of incarceration on crime by measuring the correlation across jurisdictions between incarceration rates and crime rates.

These problems are well-understood by empirical economists. A potential solution is provided by recent advances in econometrics and research design. The legal AI could run randomized control trials when law is indeterminate – that is, experiment with random decisions across different jurisdictions and then test outcomes. We already have a form of this experimentation in common law systems such as that in the United Kingdom. Judges make different decisions with some error, and then judges and researchers use this variation to learn about the effects of those decisions.

Outlook

Existing technology already allows us to use machine learning tools to predict judicial outcomes based on previous cases. A "robot clerk" based on these technologies has the potential for alleviating the overwhelming caseloads of routine decisions in many jurisdictions. Moreover, by providing some input and analysis of an individual judge's decisions, the robot clerk might help human decision-makers identify their weak spots and learn to understand and mitigate their prejudices.

But these promises do not lead to a robot serving as final legal decision-maker. The decisions of current algorithms are a black box and cannot be explained to legal participants or the public. There is no assured way to purge pre-existing bias from the machine-predicted decisions. And there are major technological and political challenges to developing and implementing a system that would read the laws and try to implement socially optimal policies.

Nevertheless, the brave new world of legal automation is upon us. These developments are both thrilling and unsettling because they attack the core of our humanity: is not justice what distinguishes man from machine? These types of questions should be the subject of democratic debate, as should the other questions raised in this report. The brief history of automated legal decision-making points to democratized, non-profit, and open-source solutions.

References

Alexander, M. (2012). The new Jim Crow: Mass incarceration in the age of colorblind-ness. The New Press.

Amaranto, D., Ash, E., Chen, D. L., Ren, L., and Roper, C. (2017). Algorithms as prosecutors: Lowering re-arrest rates without disparate impacts and identifying defendant characteristics noisy to human decision-makers.

Ash, E. and Chen, D. L. (2017). Religious freedoms, church-state separation, and religiosity: Evidence from randomly assigned judges.

Ash, E., Chen, D. L., Shang, F., Guan, X., and Yanchao, N. (2018). Judge writing style predicts decision tendencies in U.S. bankruptcy courts. Technical report.

Chen, D. L. (2014). This morning's breakfast, last night's game: Detecting extraneous factors in judging. Working paper, ETH Zurich.

Chen, D. L. and Eagel, J. (2017). Can machine learning help predict the outcome of asylum adjudications? In Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law, pages 237-240. ACM.

Danziger, S., Levav, J., and Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. Proceedings of the National Academy of Sciences, 108(17):6889-6892.

Fagan, J. and Ash, E. (2017). New policing, new segregation? from Ferguson to New York. Georgetown Law Journal.

Jurafsky, D. and Martin, J. H. (2014). Speech and language processing, volume 3. Pearson London.

Katz, D. M., Bommarito II, M. J., and Blackman, J. (2017). A general approach for predicting the behaviour of the supreme court of the united states. PloS one, 12(4):e0174698.

Keith, L. C., Holmes, J. S., and Miller, B. P. (2013). Explaining the divergence in asylum grant rates among immigration judges: An attitudinal and cognitive approach. Law 8 Policy, 35(4):261-289.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). Human decisions and machine predictions. The Quarterly Journal of Economics, 133(1):237-293.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Veiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12:2825-2830.

Posner, R. (2008). How Judges Think. Harvard University Press.

Skeem, J. L. and Lowenkamp, C. T. (2016). Risk, race, and recidivism: predictive bias and disparate impact. Criminology, 54(4):680-712.

About the author

Elliott Ash is Assistant Professor of Economics at University of Warwick, where he teaches political economy and public finance. Elliott's research focuses on empirical analysis of the law and legal system using techniques from applied microeconometrics, natural language processing, and machine learning. Elliott was previously a Postdoctoral Research Associate at Princeton University's Center for the Study of Democratic Politics. He received a PhD in economics and JD from Columbia University, a BA in economics, government, and philosophy from University of Texas at Austin, and an LLM in international criminal law from University of Amsterdam. Meanwhile, Elliott has provided expert witness testimony for the Department of Justice Civil Rights investigation into discriminatory practices at Ferguson Police Department.

About The Centre for Competitive Advantage in the Global Economy (CAGE)

Established in January 2010, CAGE is a research centre in the Department of Economics at the University of Warwick. Funded by the Economic and Social Research Council (ESRC), CAGE is carrying out a ten year programme of innovative research. The centre's research programme is focused on how countries succeed in achieving key economic objectives such as improving living standards, raising productivity, and maintaining international competitiveness, which are central to the economic wellbeing of their citizens. Its research analyses the reasons for economic outcomes both in developed economies like the UK and emerging economies such as China and India. CAGE aims to develop a better understanding of how to promote institutions and policies which are conducive to successful economic performance and endeavour to draw lessons for policy makers from economic history as well as the contemporary world. Research at CAGE examines how and why different countries achieve economic success. CAGE defines 'success' in terms of well-being as well as productivity. The research uses economic analysis to address real-world policy issues. The centre is distinctive in providing a perspective that draws on economic history as well as economic theory and is applied to countries at various different stages of economic development.

About the Social Market Foundation

The Social Market Foundation (SMF) is a non-partisan think tank. We believe that fair markets, complemented by open public services, increase prosperity and help people to live well. We conduct research and run events looking at a wide range of economic and social policy areas, focusing on economic prosperity, public services and consumer markets. The SMF is resolutely independent, and the range of backgrounds and opinions among our staff, trustees and advisory board reflects this.