DE GRUYTER Asian J Law Econ 2023; ■■■(■■■): 1–16

Research Article

Daniel L. Chen*

Incremental AI

https://doi.org/10.1515/ajle-2023-0018
Published online ■■■

Abstract: The usual narrative is backlash to artificial intelligence (AI). A recent study found that when judges were given decision-support, it ended up increasing disparities–not because the algorithm was biased–in fact the algorithm would have resulted in lower disparities. But the judges selectively paid attention to the algorithm, which resulted in greater disparities. This article argues for an incremental approach leveraging recent theoretical insights from social preference economics. The core insight is that judges are moral decision-makers–you're right or wrong, good or bad–and to understand what motivates these decision-makers, one might turn to self-image motives–a topic of active behavioral research in recent years. Each stage leverages motives related to the self: self-image, self-improvement, self-understanding, and ego. In stage 1, people use AI as a support tool, speeding up existing processes (for example, by prefilling forms). Once they're used to this, they can more easily accept an added functionality (Stage 2) in which AI becomes a choice monitor, pointing out choice inconsistencies and reminding the human of her prior choices in similar situations. Stage 3 elevates the AI to the role of a more general coach, providing outcome feedback on choices and highlighting decision patterns. Then, in stage 4, the AI brings in other people's decision histories and patterns, serving as a platform for a community of experts. This framework contrasts with the current framework where the AI simply recommends an optimal decision.

Keywords: judicial analytics, causal inference, behavioral judging

I'm going to propose how we might integrate AI incrementally into high stakes decision-making. The usual narrative is backlash to AI: when legal decision makers like judges are given decision-support, an algorithm, they might reject the AI. Indeed, in Kentucky, judges were given decision-support, an algorithm, and they found that it resulted in greater disparities - but that's not because the algorithm was

This article is based on a transcript of a presentation bringing together several streams of research.

***Corresponding author: Daniel L. Chen**, Institute for Advanced Study in Toulouse, Toulouse, 31015, France, E-mail: daniel.li.chen@gmail.com

biased (it could have been) - but the algorithm would have reduced disparities. What happened was that the judges selectively paid attention to the algorithm - so for disadvantaged groups, they ignored the algorithm, and that resulted in greater disparities (Albright 2019).

What I'm going to suggest instead is an incremental approach to integrating AI - and this is going to leverage theoretical insights from social preference economics. The core perspective is that of judges as moral decision-makers - making a decision on you're right or wrong, good or bad - and this builds of a large literature in economics on pro-social motives - Nobel lauretes Jean Tirole, Vernon Smith, Al Roth (Bénabou and Jean 2011; Roth 2007; Smith 2016) – all talking about why people might make pro-social decisions and thinking self-image ("I think I'm a good person - a good judge") might be important.

As a preliminary step, let's go back to an earlier AI literature on decision-making, which is to compare a decision-maker to the predicted self. It's called bootstrapping (Camerer 1981). The idea is that we might ask, are humans actually making more errors than their predicted self or vice versa. Let's suppose I'm a very attentive judge, paying attention to a lot of legally relevant factors that aren't incorporated in the machine learning model, then you might expect the human to do better, to make fewer errors, to be having less disparate impacts than the bootstrapped judge. But if instead, I'm human, and I pay attention to extraneous factors, I may actually end up making more mistakes relative to my predicted self (Kleinberg et al. 2018).

Once we first verify that it's possible that a bootstrapped judge might do better than a human judge – which can be assessed simply with historical data, we might introduce the judge to their predicted self, using AI as a support tool, speed up existing processes. Think of prefilling forms, or on our smartphones, we have autocomplete when texting people. So, what might showing the predicted self do? First, it leverages self-image motives, I get to see what I would have done. Second, it cabins extraneous factors, making us more consistent comparing apples and apples, oranges and oranges. Third, it presents a default that I can easily deviate from, so it might leverage more conscious deliberate Type II thinking as opposed to Type I thinking.

- *Stage 1 has people using AI as a support tool, speeding up existing processes (for example, by prefilling forms). Offer a judge the best prediction of themselves, based on the specific judge's previous decision-making, from a model using only legally relevant features.*

In Stage 2, once judges are used to this, they might more easily accept added functionality, like a light nudge, choice monitor, pointing out when they're inconsistent with their previous decisions. Let's suppose the algorithm predicts that when you deviate from the default, you're more likely to make an error, to be more mistaken, to be reversed. The light nudge might be to tell the judge to pay more attention, be less indifferent, spend more time, and collect more information on this case. This might leverage motives of self-improvement to be less erroneous, to be a better judge.

- *Stage 2 has **prediction of reversal and the motive to self-improve, while also respecting judge's autonomy and limited bandwidth.***

Now we can go to Stage 3, where we think of AI as a more general coach that explains why decisions might lead to error. Let's suppose judges want transparent AI. Research suggests that transparency can makes AI more trustworthy, where trust comes from causal understanding (Danks 2019). One might provide a dropdown menu so judges can understand why the algorithm is nudging him or her. The judge gets to find out, here's the reason why s/he's historically made more errors with this particular deviation from the default. This might leverage motives of self-understanding. But it doesn't have to be a one-way conversation, where the AI is talking to the judge. The judge can also explain to the AI why I am choosing to deviate, so the AI can learn from the judge, to address the issue of status quo bias, where the world is changing.

- *Stage 3 elevates the AI to the role of a more general coach, providing outcome feedback on choices and highlighting decision patterns. The judge can also explain the reason for deviating.*

Then we turn to Stage 4, where you might bring in other people's decision histories. If we can show the judge the predicted self, why not show the predicted others, so the judge has a platform to access a community of experts. This leverages many of the previous motives related to the self but also motives related to comparisons with others.

- *Stage 4 has the AI bring in other people's decision histories and patterns, serving as a platform for a community of experts.*

Only in Stage 5, you might recommend the optimal decision if there is one.

The theory behind the piece is one of ego or self-concept, an idea that can be traced to Enlightment and Romantic ideals of the self—Enlightenment ideals of self-knowledge and self-mastery, presuming that each person has an interior space and memory of oneself, and Romantic ideals of self-realization, distinctive self-expression and authenticity, being true to oneself (Taylor 1989, 1991). Recognizing that there are different sources of the self, according to philosophers, constituted a moral revolution, in which projects of personal identity (a self constituted by one's strongest values and commitments) came to be important beyond economic self-interest. Questions of identity and self-fulfilment became personally significant; refusals of acceptance and respect, deeply challenging. We can see projects of personal identity being mobilized in both progressive (gender and sexual identities and claims to equal rights) and conservative politics (national and communal identities defended by populist movements). We can see projects of personal identity in self-esteem and authenticity being taught as values in school—"find your passion". We can see projects of personal identity and self-esteem—violations of self-esteem—in contemporary discussions of microaggression, trigger warnings, and privilege disparities. Recognizing that everyone has their own way of being human—especially for those whose identities have been systematically degraded and whose rights to be treated as equals have been neglected—facilitates respect for individuals, but also for different cultures, according to philosophers.

Now let me make this more concrete. Much of my research thinks about concepts of legitimacy in law (Chen 2013). Coming from economics, a lot of theorists will write down models where people are motivated by law through sanctions and deterrence. But many non-economists will argue that sometimes law's just legitimate - that people might follow the law because it's the right thing to do, and make their decision for reasons apart from the consequences of decisions. But this motive might presume that the law or the lawmaker is legitimate – that the lawmaker is not indifferent, and recognizes my individuality. So that might explain some of the backlash to AI, because an algorithm treats me like a datapoint like all the others, but I think of myself as unique, and a little more different; I want the judge to recognize my specific circumstances.

Here's an anecdote that captures one way to think about indifference, where police officers in the U.S. South literally flipped a virtual coin before making an arrest, a potentially life-changing decision.[1] So why might this violate our notions of justice and what does machine learning have to do with this?

We can think of two definitions of justice - equal treatment before the law and equality based on recognition of difference. So think of a set of features X that should lead to the same predictions or predictive accuracy of outcomes, and a set of W's that should not. We tend to think of X's as consequences of our actions, based on of the control principle and merit principle that we should be liable only for events or actions that are under our control (Gurdal et al. 2013; Moulin 2004). But then the W's might be things that I think many justice systems will recognize and think should be orthogonal to outcomes, like ethnicity, gender, masculinity, or name. The W's might also be things outside of our control - like football, weather, judge's lunchtime, or preceding case. So think of the X's as recognizing our individuality, that should lead to the same predictions, and then the W's are things that we, as a justice system, has said should be orthogonal to outcomes, and when these last group of features are affecting decisions, that's going to affect the predictability of outcomes.

Behavioral anomalies offer an intuitive understanding of feature relevance. Psychologists have documented many effects of moderate size in the lab, so "settings where people are closer to indifference between options are more likely to lead to detectable effects of behavioral biases outside the lab. So suppose a parabola

[1] https://www.nytimes.com/2018/07/14/us/police-coin-flip-arrest.html.

captures what the judge thinks it the right decision, small deviations cost little, the judge isn't sent to jail. But now suppose the parabola flattens – 8 months, 8 years – I'm not sure, I'm inattentive – or – I don't care, I'm indifferent, then as the influence of the legally relevant factors wanes, the influence of extraneous factors can grow, so we can think of measuring behavioral bias as documenting what economists might call "revealed preference indifference".

(I) *Using machine learning to detect judicial indifference to case features*

Machine learning can help diagnose judicial inattention in at least five different ways. First is early predictability. Think about predicting the decision when the case opens versus when it closes. It turns out in the U.S., you can predict with roughly 80 % accuracy when the case opens for immigration judges versus when it closes. So this might indicate circumstances where things during the case hearings are not really affecting the final outcomes of the cases, where they might be using more snap judgments or heuristics.

Dunn et al. (2017) conceptualizes early predictability, the possibility to use machine learning to automate the detection of judicial indifference, where the judges appear to ignore the circumstances of the case, *X*. In asylum courts, judges can be predicted with the same level of accuracy at the time the case opens — and at the time the case closes. To be sure, there may be external circumstances like country war that should dictate the outcome of the case. But significant inter-judge disparities in predictability suggest not all the judges interpret that information in the same way, raising questions of snap or predetermined judgment. Figure 1 shows that judges with low and high grant rates are more predictable.

(II) *Using machine learning to automate the detection of judicial inconsistencies*

A second way to diagnose judicial inattention is the detection of behavioral anomalies - like the ones I've been describing - whether you do this through a larger set of features for machine learning or do causal inference. Let me demonstrate this idea with the asylum courts where I have the administrative universe since 1981. This data comprise half a million asylum decisions across 336 hearing locations and 441 judges. The applicant for asylum reasonably fears imprisonment, torture, or death if forced to return to their home country. The average grant rate is about 35 %. Applicants are randomly assigned. Dunn et al. (2017) shows that using data only available up to the decision date, you can achieve 80 % predictive accuracy. Chen and Eagel (2017)Chen 2017 shows this is predominately driven by trend features and judge characteristics, *W*'s beyond the applicant control, that might raise questions of due process violations. About one-third is driven by case information, news events, and court information.

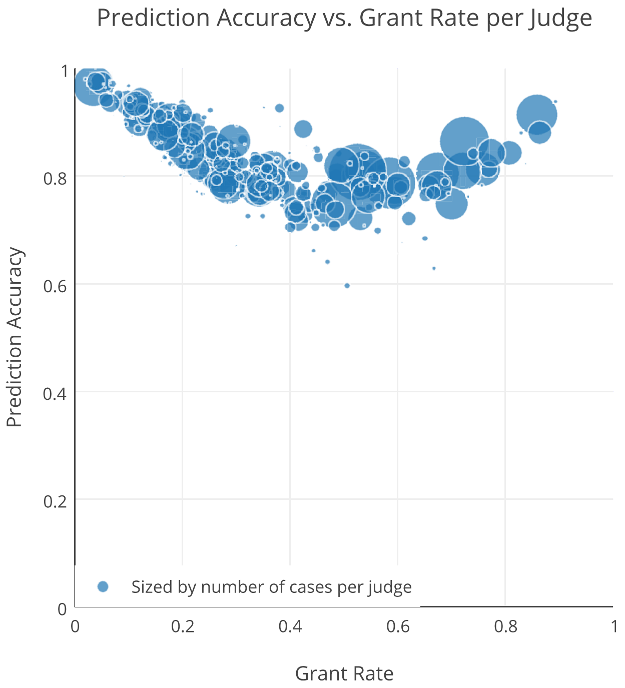


Figure 1: Early predictability of asylum decisions. Judges with high and low grant rates are more predictable Source: Dunn et al. (2017).

Figure 2 shows some descriptive statistics. Judges are more lenient before lunch and towards the end of the day. The lower left of Figure 2 shows that there is a U-shape relationship with family size, and the lower right shows that defensive cases are less likely to be granted – defensive cases are those where the applicant has been caught, rather than applying for an extension to stay. Figure 3 shows that judges are more lenient with good weather rather than extreme weather and more lenient with a genocide news indicator. The bottom part shows strong trend factors both within the court on the left and over time on the right. While the literature typically studies one behavioral feature at a time, Chen and Eagel (2017)Chen 2017 demonstrates the possibility for machine learning to automate the detection of judicial inconsistencies due to W.

(III) *Using machine learning to directly detect indifference to appeals.*

Now, suppose judges are systematically indifferent, so they might be inattentative to reversals, especially surprise reversals - the decisions predicted to be affirmed. So, machine learning for predicting decisions when cases open, predicting decisions

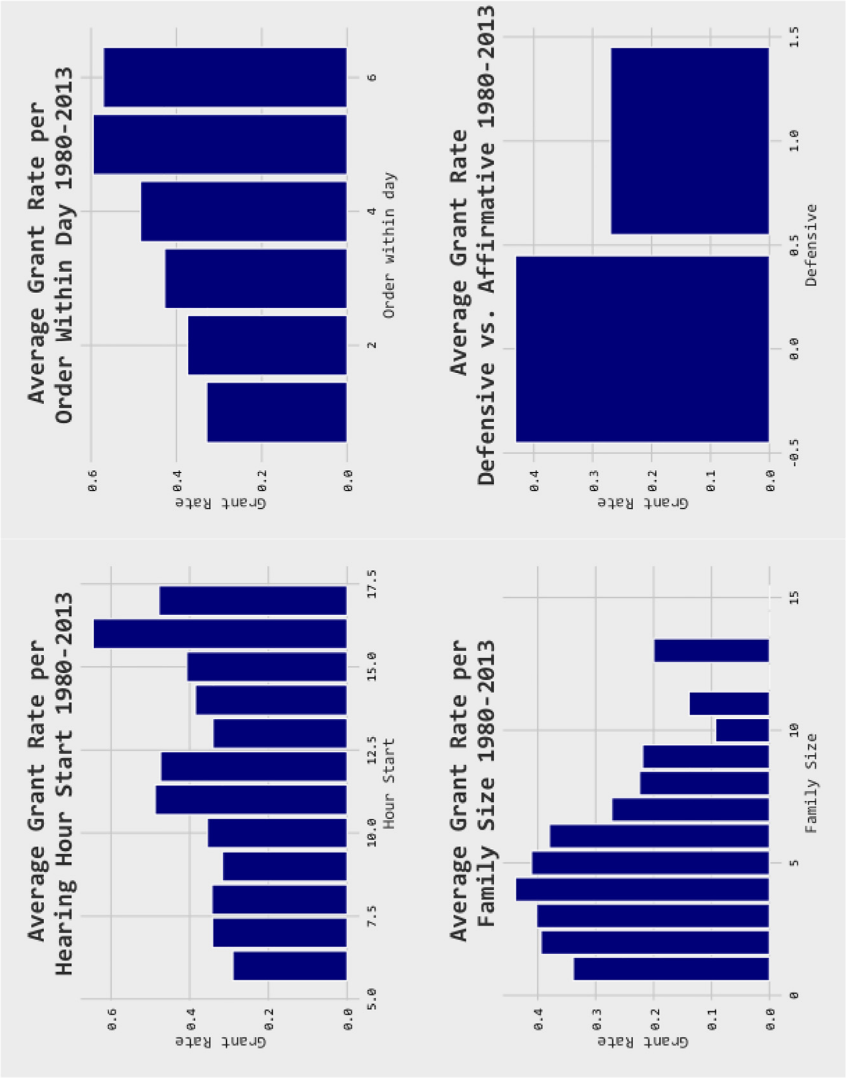


Figure 2: Predictability of asylum Decisions. More lenient before lunch & towards end of day & for affirmative asylum, U-shape with family size Source: Chen and Eagel (2017).

when cases close, having a larger set of features to do predictions, and then predicting the reversal. Figure 4 shows that asylum judges respond to surprise reverals by granting more asylum and holding more hearing sessions.

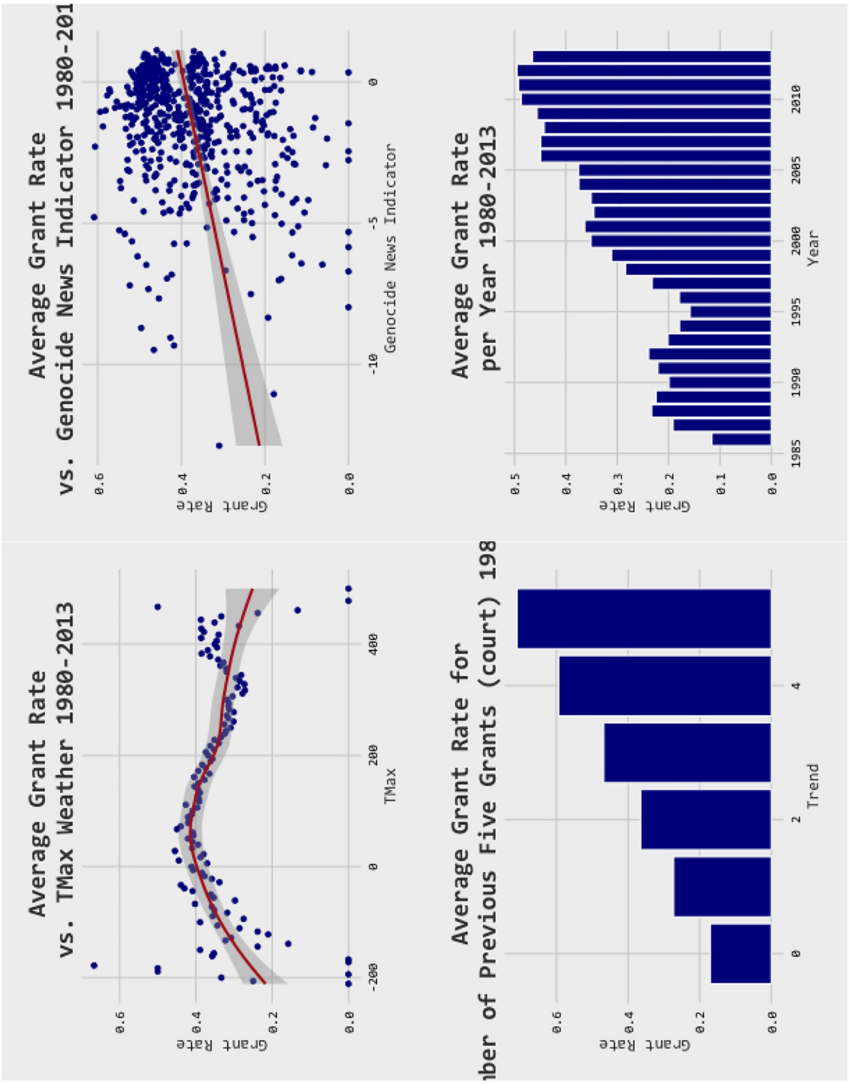


Figure 3: Predictability of asylum decisions. More lenient with good weather & genocide news indicator & over time & female judges. Strong trend factors within-court & within-judge. Source: Chen and Eagel (2017).

(IV) *Do less attentive judges have implicit risk rankings closer to random?*

Judges vary substantially in their responsiveness to reversal. We can use machine learning to assess whether less attentive judges are closer to flipping a coin, such that

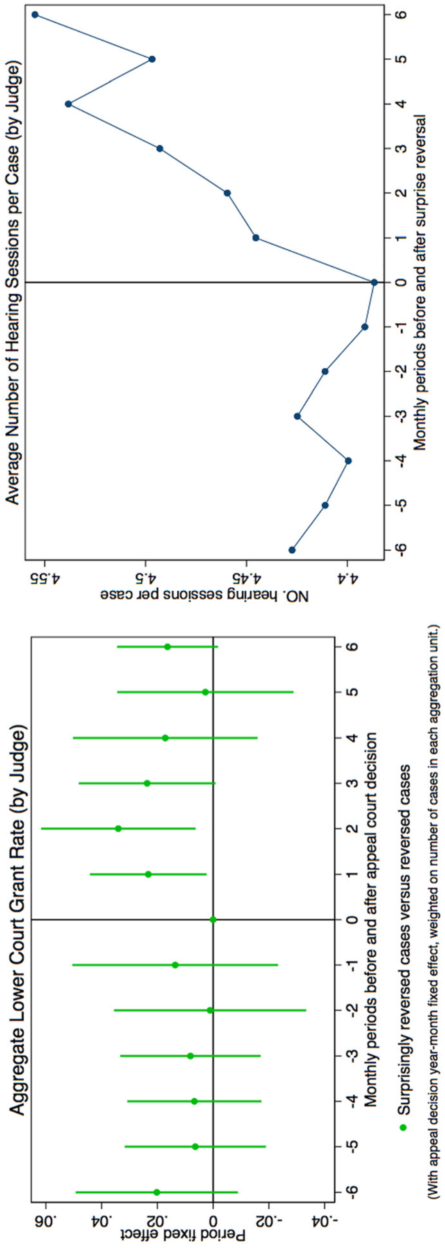


Figure 4: Attentativeness to surprise reversals. Source: Ash and Chen (2020).

their implicit risk rankings are closer to random. We don't have the actual risk ranking of each litigant but we can assess the implicit risk ranking by comparing the stricter and more lenient decision-makers.

Consider the following hypothetical in the prosecutorial decision to screen an individual into the criminal justice system or to send the individual home. Amaranto et al. (2017) use data from New Orleans for 12 years. The data is incredibly detailed - 430,000 charges, 145,000 defendants - and a 594 page codebook including the name, social security number, victims, witnesses, police officers, and so on–a perfect setting for machine learning, with more columns than rows. The entire data collection process begins at the moment of the arrest to the final sentence, if any (Figure 5).

To put this in perspective, other datasets are not linked: they study victimization, or police reports, or arrests, and the many random judge assignment studies examine (and can only examine) the final node only. This data reveals the screening (rejections and dismissals) decision. Why is that important? Amaranto et al. (2017) addresses the broader disparities in criminal justice, where motivations like the perceived legitimacy of the lawmaker have been hotly debated alongside racial differences in the police use of force.

An unexamined issue is what happens after an arrest and before the trial. Information about cases dropped by the prosecutor has been-to date-unavailable—and they are largely unaccountable. The prosecutor is said to decide the fate of 15 cases for every case presided over in trial (Wilson and Petersilia 2010). From 1990 to

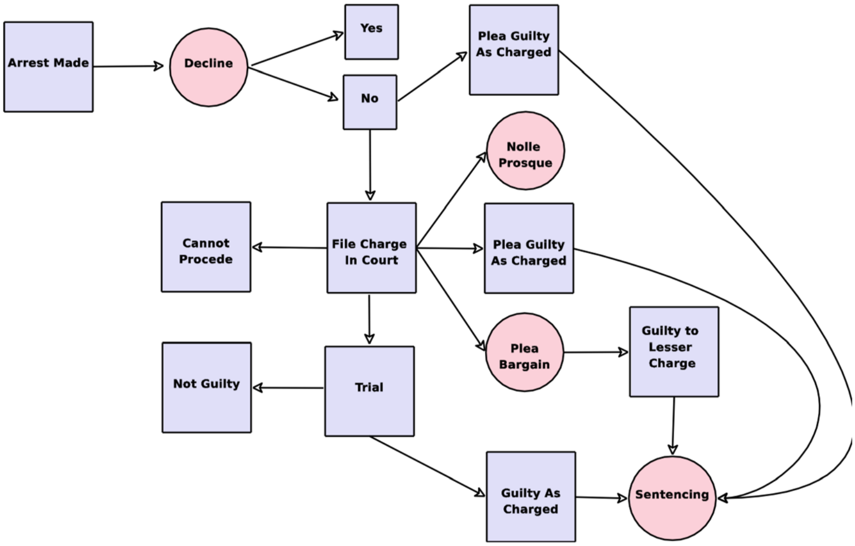
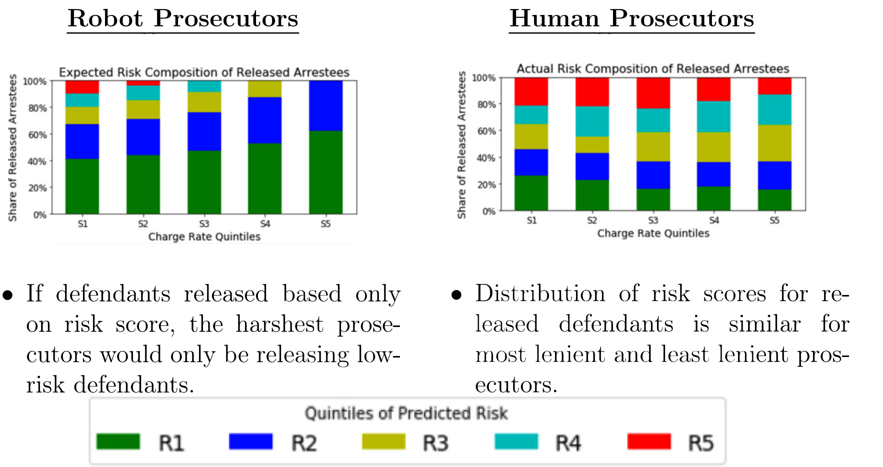


Figure 5: Detailed information on arrests linked to final resolution

2010, roughly 50 % of increase in felony filings comes from misdemeanors being charged as felonies (Pfaff 2017). A recent study argues that the prosecutor charge type can essentially make the racial gap in sentences disappear (Rehavi and Starr 2014). Prosecutors are therefore very powerful. The data shows the racial gap reappears once prosecutor screening is taken into account. Namely, if Black defendants are less likely to have their case screened out, that mechanically reintroduces the sentencing gap in a very real manner. Cases are randomly or rotationally assigned to screeners. The screening decision is interpretive and discretionary.

How the screeners rank the risk of the arrestees is unobserved. However, we can assess their implicit risk ranking by comparing the distribution of predicted risk of the arrestees charged by the "strict" and the "lenient" screeners. That is, we can use machine learning to understand how screeners screen. The actual risk distribution amongst strict and lenient screeners differ from what we would expect to see if the screeners were releasing based on predicted risk.

If a defendant were released based only on risk score, the harshest quintile of prosecutors, represented by S5 in Figure 6, would be releasing only a few individuals and we might expect only the low-risk R1s and R2s to be released. As we move towards the most lenient prosecutors, represented by S1, we would start to see R4s and R5s. If instead we were flipping a coin, we would expect the same distribution of risk scores across the strict and lenient prosecutors and within each quintile, an even distribution of R1 to R5 within each bar. This is what we actually see with human prosecutors on the right.



Robot Prosecutors

Human Prosecutors

- If defendants released based only on risk score, the harshest prosecutors would only be releasing low-risk defendants.

- Distribution of risk scores for released defendants is similar for most lenient and least lenient prosecutors.

Figure 6: Implicit risk rankings. Source: Amaranto et al. (2017).

Turning to asylum, we can ask whether the lenient asylum judges are only denying the riskiest applicants, the ones who would have their appeal rejected–that is, we should see the lowest reversal rates of their asylum denials. However, if I am a harsh asylum judge, rejecting most of my applicants, I would expect a larger share of my denials be overturned. What we see on the left side of Figure 7 is that judges have strong habits: a judge who is generally lenient in other cases is likely to be lenient in any given case. In the right side, the judges are sorted from least to most lenient. If judges are 'ordering' their asylees, the most lenient judge letting in the most applicants should be rejecting only the "least safe" applicants, so we should expect the appeal success to be lower. This is what we see among the more attentive judges but not the less attentive judges, who may be more prone to other extraneous factors such as the weather, as we see in Figure 8.

(V) *Using machine learning to detect difference-in-indifference*

And then we might investigate difference-in-indifference. So for example is inattentiveness exacerbated for those from the Global South, which turns out to be the case as shown in Figure 9. Likewise, difference-in-indifference can be investigated for early predictability, mistakes, and behavioral anomalies.

These five uses of machine learning are all a bit negative and diagnostic. But we can also use machine learning to increase recognition and dignity.

So now let me talk about how might this be done in practice, this decision-support. An app can be given to a judge: here's the schedule type, judge name, a dropdown menu to access themselves or the others, the city, the location, the time, etc., so they can view the prediction. Then, we might do an impact evaluation on trust, perceived indifference of lawmakers, but also more pragmatic things like the number of applications, how the decisions evolve, reversals, the speed, and the disparities.

At a less granular level, court administrators and judges can already access their court statistics – what is it I have done in the past, how other tribunals have done it, how the other judges are doing it. We can test the role of self-image motivations. Suppose the program has diagnosed that you are slow relative to others, or that you are more absent relative to others. And, these are life-tenured judges with fixed financial incentives, so you are more absent than your colleagues, you're slower than your colleagues - you pick the index that the historical data has already shown you to be worst performing on.

Judicial analytics holds the promise of increasing efficiency and fairness of law. While much empirical work evaluates judges to observe inconsistencies in their behavior, the advent of machine learning tools offers an approach to automate the detection of inconsistencies and inattention. We are working with various country governments to impact evaluate personalized nudges for judges, telework with non-financial based congestion pricing, mobile/e-justice apps, GPS-based scheduling of

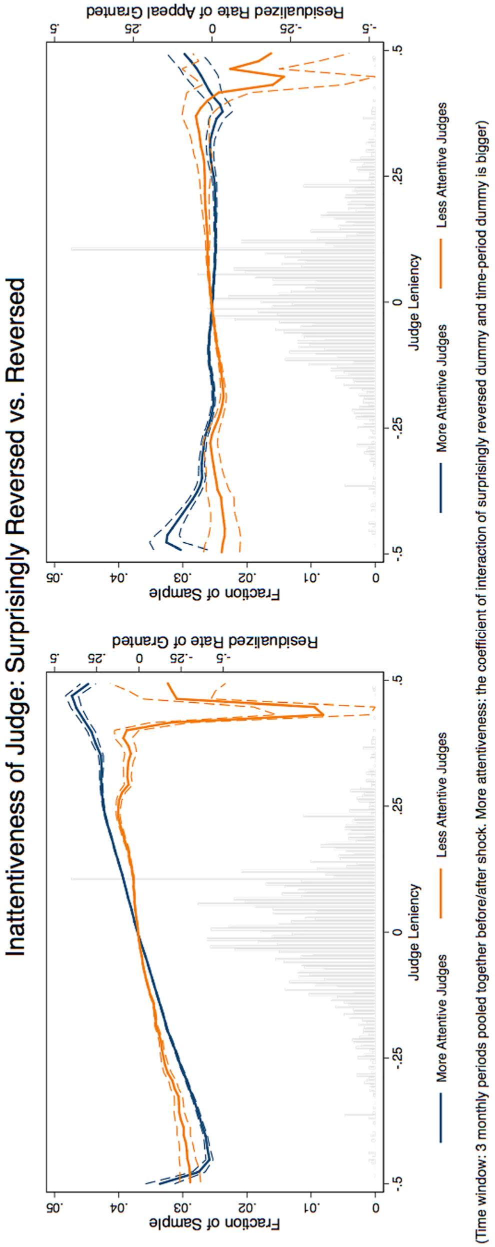


Figure 7: Implicit risk rankings. Source: Ash and Chen (2020).

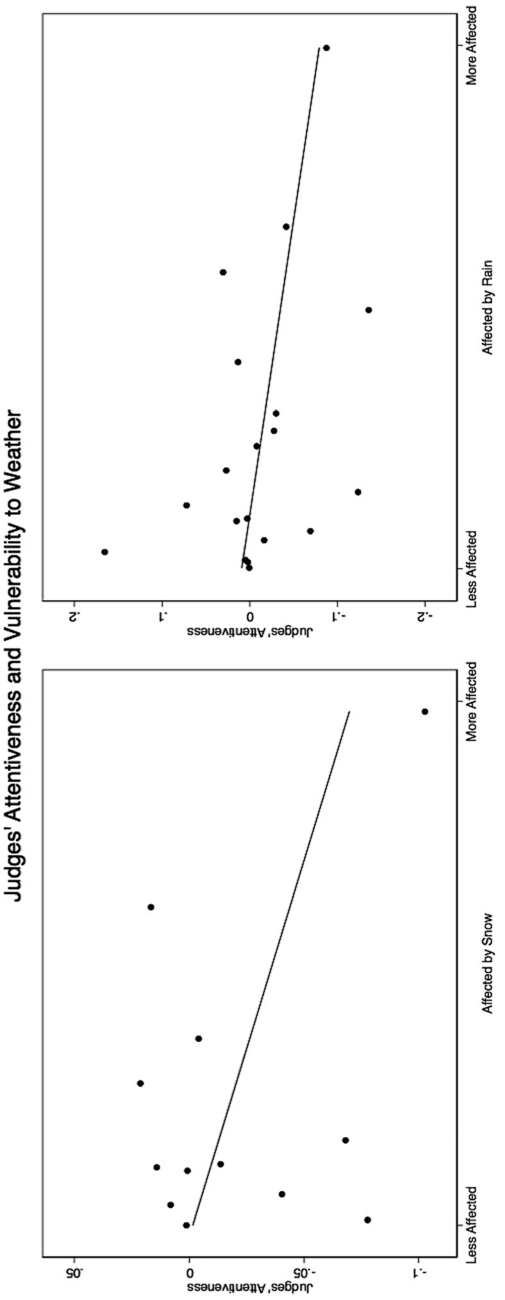


Figure 8: Attentiveness and proneness to extraneous factors. Source: Ash and Chen (2020).

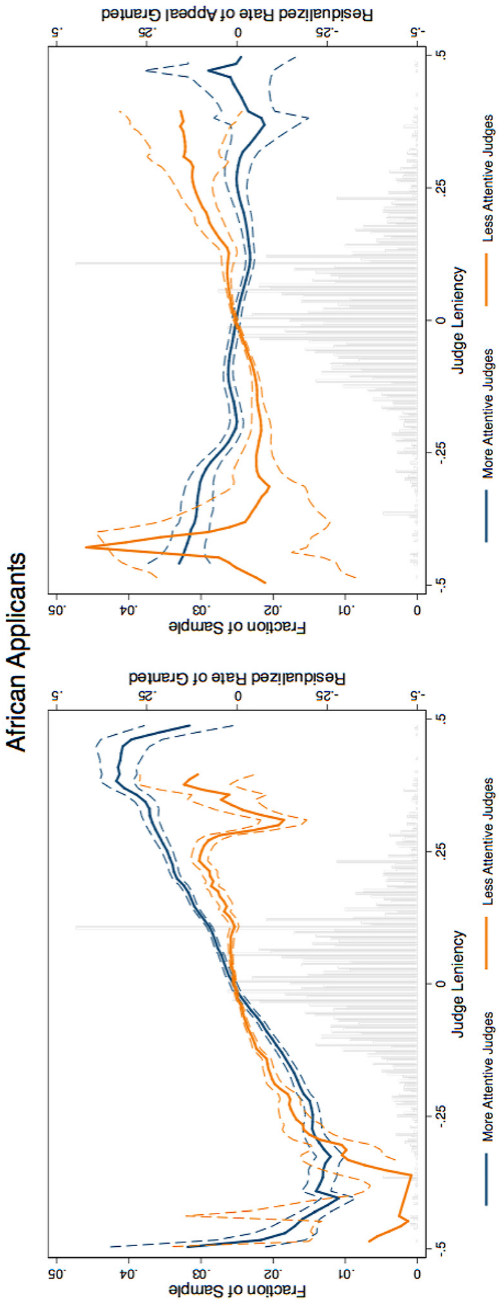


Figure 9: Implicit risk rankings. Source: Ash and Chen (2020).

hearings, adaptive learning, theory versus case-based teaching (using the history of their own past decisions), social-emotional learning (self-reflection) interventions, legal analytics, and other fairness, accountability, and transparency initiatives around AI and justice.

References

Albright, A. 2019. “If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions.” In *The John M. Olin Center for Law, Economics, and Business Fellows’ Discussion Paper Series*, 85. Cambridge, MA: Harvard Law School.

Amaranto, D., A. Elliott, D. L. Chen, L. Ren, and C. Roper. 2017. *Algorithms As Prosecutors: Lowering Rearrest Rates without Disparate Impacts and Identifying Defendant Characteristics Ânoisyâto Human Decision-Makers*.

Ash, E., and D. Chen. 2020. *Judicial Inattention: Machine Prediction of Appeal Success in U.S. Asylum Courts*.

Bénabou, R., and T. Jean. 2011. “Identity, Morals, and Taboos: Beliefs as Assets.” *The Quarterly Journal of Economics* 126 (2): 805–55. http://www.jstor.org/stable/23015689. ISSN 00335533.

Camerer, C. 1981. “General Conditions for the Success of Bootstrapping Models.” *Organizational Behavior & Human Performance* 27 (3): 411–22.

Chen, D. L. 2013. “The Deterrent Effect of the Death Penalty? Evidence From British Commutations during World War I.” Working Paper. ETH Zurich. Also available at http://nber.org/~dlchen/papers/DeathPenalty.pdf.

Chen, D.L. and J. Eagel. 2017. “Can Machine Learning Help Predict the Outcome of Asylum Adjudications?” In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, 237–40.

Danks, D. 2019. “The Value of Trustworthy Ai.” In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 521–522.

Dunn, M., L. Sagun, H. Şirin, and D. Chen. 2017. “Early Predictability of Asylum Court Decisions.” In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, 233–6.

Gurdal, M. Y., J. B. Miller, and A. Rustichini. 2013. “Why Blame?” *Journal of Political Economy* 121 (6): 1205–47.

Kleinberg, J., H. Lakkaraju, J. Leskovec, and J. Ludwig. 2018. “Human Decisions and Machine Predictions.” *The Quarterly Journal of Economics* 133 (1): 237–93.

Moulin, H. 2004. *Fair Division and Collective Welfare*. Cambridge, MA: MIT press.

Pfaff, J. F. 2017. “Locked.” In *The True Causes of Mass Incarceration and How to Achieve Real Reform*. New York, NY: Basic Books.

Rehavi, M. M., and S. B. Starr. 2014. “Racial Disparity in Federal Criminal Sentences.” *Journal of Political Economy* 122 (6): 1320–54.

Roth, A. E. 2007. “Repugnance as a Constraint on Markets.” *The Journal of Economic Perspectives* 21 (3): 37–58.

Smith, V. L. 2016. “The Fair and Impartial Spectator.” *Econ Journal Watch* 13 (2): 330–9.

Taylor, C. 1989. *Sources of the Self: The Making of the Modern Identity*. Cambridge, MA: Harvard University Press.

Taylor, C. 1991. *The Ethics of Authenticity*. Cambridge, MA: Harvard University Press.

Wilson, Q. J., and J. Petersilia. 2010. *Crime And Public Policy*. New York, NY: Oxford University Press.