University of Maryland Francis King Carey School of Law
DigitalCommons@UM Carey Law

Faculty Scholarship | Francis King Carey School of Law Faculty

2018

Prediction, Persuasion, and the Jurisprudence of Behaviorism

Frank A. Pasquale
University of Maryland Francis King Carey School of Law, fpasquale@law.umaryland.edu

Glyn Cashwell

Follow this and additional works at: https://digitalcommons.law.umaryland.edu/fac_pubs

Part of the Internet Law Commons, Law and Society Commons, and the Science and Technology Law Commons

Digital Commons Citation
Pasquale, Frank A. and Cashwell, Glyn, "Prediction, Persuasion, and the Jurisprudence of Behaviorism" (2018). *Faculty Scholarship*. 1604.
https://digitalcommons.law.umaryland.edu/fac_pubs/1604

This Article is brought to you for free and open access by the Francis King Carey School of Law Faculty at DigitalCommons@UM Carey Law. It has been accepted for inclusion in Faculty Scholarship by an authorized administrator of DigitalCommons@UM Carey Law. For more information, please contact smccarty@law.umaryland.edu.

Prediction, Persuasion, and the Jurisprudence of Behaviorism

Frank Pasquale and Glyn Cashwell[1]

A growing chorus of critics has challenged the use of opaque (or merely complex) predictive analytics programs to monitor, influence, and assess individuals' behavior.[2] The rise of a "black box society" portends profound threats to individual autonomy: when critical data and algorithms cannot be a matter of public understanding or debate, both consumers and citizens are unable to comprehend how they are being sorted, categorized, and influenced.[3]

A predictable counterargument has arisen, discounting the comparative competence of human decision-makers. Defending opaque sentencing algorithms, for instance, Christine Remington (a Wisconsin assistant attorney general) stated, "We don't know what's going on in a judge's head; it's a black box, too."[4] Of course, a judge must (upon issuing an important decision) explain why the decision was made; so too are agencies covered by the Administrative Procedure Act obliged to offer a "concise statement of basis and purpose" for rulemaking.[5] But there is a long tradition of realist commentators dismissing the legal justifications adopted by judges and administrators as unconvincing fig leaves for the "real" (non-legal) bases of their decisions.

In the first half of the twentieth century, the realist disdain for stated rationales for decisions led in at least two directions: toward more rigorous and open discussions of policy considerations motivating judgments, and toward frank recognition of judges as political actors, reflecting certain ideologies, values, and interests. In the 21st century, a new response is beginning to emerge: a deployment of natural language processing (NLP) and machine learning (ML) techniques to predict whether judges will hear a case, and if so, how they will decide it. Machine learning experts are busily feeding algorithms the opinions of the US Supreme Court, European Court of Human Rights, and other judicial bodies, as well as metadata on justices' ideological commitments, past voting record, and myriad other variables. By processing data related to cases, and the text of opinions, these systems purport to predict how judges will decide cases, how individual judges will vote, and how to optimize submissions and arguments before them.

This form of prediction is analogous to forecasters using big data (rather than understanding underlying atmospheric dynamics) to predict the movement of storms. An algorithmic analysis of a database of, say, 10,000 past cumulonimbi sweeping over Lake Ontario, may prove a better predictor of the next cumulonimbus's track, than a trained meteorologist without access to such a data trove. From the perspective of many predictive analytics approaches, judges are just like any other living feature of the natural world - an entity that transforms certain inputs (such as briefs and advocacy documents) into outputs (decisions for or against a litigant). Just as forecasters predict whether a cloud will veer southwest or southeast, the user of a machine learning system might use machine-readable case characteristics to predict whether a rainmaker will prevail in the courtroom.

We call the use of algorithmic predictive analytics in judicial contexts an emerging jurisprudence of behaviorism, as it rests on a fundamentally Skinnerian model of cognition as a black boxed transformation of inputs into outputs.[6] In this model, persuasion is passé; what matters is prediction.[7] After describing and critiquing a recent study that has advanced this jurisprudence of behaviorism, we question the value of such research.

Billed as a method of enhancing the legitimacy and efficiency of the legal system, such modeling is all too likely to become one more tool deployed by richer litigants to gain advantages over poorer ones.[8] Moreover, it should raise suspicions if it is used as a triage tool to determine the priority of cases. Such predictive analytics are also only as good as the training data they depend on. While fundamental physical laws rarely if ever change, human behavior can change dramatically in a short period of time. Therefore, one should always be cautious when applying automated methods in the human context, where factors as basic as free will and political change make the behavior of both decision-makers, and those they impact, impossible to predict with certainty.[9]

Nor are predictive analytics immune from bias. Just as judges bring biases into the courtroom, algorithm developers are prone to incorporate their own prejudices and priors into their machinery.[10] Nor are biases easier to address in software than in decisions justified by natural language. Such judicial opinions (or even oral statements) are generally much less opaque than machine learning algorithms. Unlike many proprietary or hopelessly opaque

computational processes proposed to replace them, judges and clerks can be questioned and rebuked for discriminatory behavior.[11]

There is a growing literature critiquing the unreflective application of machine learning techniques to social problems.[12] Predictive analytics may reflect biases rather than reasoned decisionmaking.[13] They also leave those affected by automated sorting and categorizing unable to understand the basis of the decisions affecting them, assuming that the output from the models in anyway affects one's life, liberty, or property rights and that litigants are not given the basis of the model's predictions.[14] This article questions the social utility of prediction models as applied to the judicial system, arguing that their deployment may endanger core rule of law values. In full bloom, predictive analytics would not simply be a camera trained on the judicial system, reporting on it, but it would also be an engine of influence. Attorneys may decide whether to pursue cases based on such systems; courts swamped by appeals or applications may be tempted to use machine learning models to triage or prioritize cases.

In work published to widespread acclaim in 2016, Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos made bold claims about the place of natural language processing in the legal system in their article, *Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective* ("*Predicting Judicial Decisions*").[15] They claim that "advances in Natural Language Processing (NLP) and Machine Learning (ML) provide us with the tools to automatically analyse legal materials, so as to build successful predictive models of judicial outcomes."[16] Presumably they are referring to their own work as part of these advances. However, close analysis of their "systematic study on predicting the outcome of cases tried by the European Court of Human Rights based solely on textual content" reveals that their *soi-disant* "success" merits closer scrutiny, on both positive and normative grounds.

The first questions to be asked about a study like *Predicting Judicial Decisions* are: what are its uses and purposes? Aletras et al. suggest at least three. First, they present their work as a first step toward the development of machine learning (ML) and natural language processing (NLP) software that can predict how judges and other authorities will decide legal disputes. Second, Nikos Aletras has clearly stated to media that artificial intelligence "could also be a valuable tool for highlighting which cases are most likely to be violations of the European

Convention of Human Rights"—in other words, that it could help courts triage which cases they should hear.[17] Third, they purport to intervene in a classic jurisprudential debate—whether facts or law matter more in judicial determinations.[18] Each of these aims and claims should be rigorously interrogated, given shortcomings of the study that the authors acknowledge. Beyond those acknowledged problems, there are even more faults in their approach which cast doubt on whether the research program of NLP-based prediction of judicial outcomes, even if pursued in a more realistic manner, has anything significant to contribute to our understanding of the legal system.

Though Aletras et al. have used rigorous ML and NLP methods in their study, their approach metaphorically stacks the deck in so many ways that it is hard to see its relevance to either practicing lawyers or scholars. Nor is it likely to be ethical for attorneys to use even refined approaches arising out of this line of research, to influence judges by deploying words or word structure found likely to have stimulated positive outcomes in the past, if such a deployment of the method is premised on subliminal or otherwise undetectable influence. Nor is it plausible to state that a method this crude, and disconnected from actual legal meaning, provides empirical data relevant to jurisprudential debates over legal formalism and realism. There is an inevitable interface of human meaning that is necessary to make sense of social institutions like law. Machine "learning" is nowhere near simulating it, let alone understanding it in recognizably human terms.[19]

II. Stacking the Deck: "Predicting" the Contemporaneous

The European Court of Human Rights (ECHR) hears cases in which parties allege that their rights under Articles of the European Convention of Human Rights (Convention) were violated and not remedied by their country's courts.[20] The researchers claim that the textual model has an accuracy of "79% on average."[21] Given sweepingly futuristic headlines generated by the study (including "Could AI replace judges and lawyers?"),[22] a casual reader of reports on the study might assume that this finding means that, using the method of the researchers, those who have some aggregation of data and text about case filings, can use that data to predict how the ECHR will decide a case, with 79% accuracy. But that would not be accurate. Instead, the researchers used the "'Circumstances' subsection" in the cases they claimed to "predict," which had "been formulated by the Court itself."[23] In other words, they claimed to be "predicting" an

event (a decision) based on materials released simultaneously with the decision. This is a bit like claiming to "predict" whether a judge had cereal for breakfast yesterday, based on a report of the nutritional composition of the materials on the judge's plate at the exact time she or he consumed the breakfast.[24] Readers can (and should) balk at using the term "prediction" to describe correlations between past events (like decisions of a court) and contemporaneously generated, past data (like the Circumstances subsection of a case). Sadly, though, few journalists breathlessly reporting the Aletras et al. study did so—and the authors do not appear to have done enough to disabuse them of their sensational extrapolations.

To their credit, Aletras et al. repeatedly emphasize how much they have effectively stacked the deck by using *ECHR-generated documents themselves* to help the ML/NLP software they are using in the study "predict" the outcomes of the cases associated with those documents. A truly predictive system would use filings of the parties, or data outside the filings, in existence before the judgement itself. Aletras et al. grudgingly acknowledge that the Circumstances subsection "should not always be understood as a neutral mirroring of the factual background of the case," but defend their method by stating "summaries of facts found in the 'Circumstances' section have to be at least framed in as neutral and impartial a way as possible."[25] However, they give readers no clear guide as to *when* the Circumstances subsection is actually a neutral mirroring of factual background, or how closely it relates to records in existence before a judgment that would actually be useful to those aspiring to develop a predictive system.

Instead, their "premise is that published judgments can be used to test the possibility of a text-based analysis for ex ante predictions of outcomes *on the assumption that* there is enough similarity between (at least) certain chunks of the text of published judgments and applications lodged with the Court and/or briefs submitted by parties with respect to pending cases."[26] But they give us few compelling reasons to accept that assumption, since almost any judge writing an opinion to justify a judgment is going to develop a facts section fully aware of how its preferred interpretation of the law is more likely to be supported by some renditions of the facts rather than others. The authors state that the ECHR has "limited fact finding powers," but give no sense of how much that mitigates that problem of law-driven cherry-picking (highlighting favorable facts) or lemon-dropping (failing to mention unfavorable facts). Nor should we be comforted (as the authors apparently are) by the fact that "the Court cannot openly acknowledge any kind of bias

on its part"—indeed, that suggests a need for the Court to avoid the main type of transparency in published justification that could help researchers artificially limited to NLP better understand it.[27] The authors also state that in the "vast majority of cases," the "parties do not seem to dispute the facts themselves, as contained in the 'Circumstances' subsection, but only their legal significance." However, the critical issues here are, first, the facts themselves, and second, how the parties characterized the facts *before* the Circumstances section was written. Again, the fundamental problem of mischaracterization—of "prediction," instead of mere correlation or relationship—crops up to undermine the value of the study.

Even in its most academic mode—as an ostensibly empirical analysis of the prevalence of legal realism—the Aletras et al. study stacks the deck in its favor in important ways. Indeed, it might be seen as assuming at the outset a version of the very hypothesis it ostensibly supports. That hypothesis is that something other than legal reasoning itself drives judicial decisions. Of course that is true in a trivial sense—there is no case if there are no facts—and perhaps the authors intend to make that trivial point.[28] But their language suggests a larger aim, designed to meld NLP and jurisprudence. Given the critical role of meaning in the latter discipline, and their NLP methods' indifference to it, one might expect an unhappy coupling here. And that is indeed what we find.

In Aletras et al.'s study, the corpus used for the predictive algorithm was a body of Court of Human Rights' "published judgments." Within these judgments, a summary of the factual background of the case was summarized (by the Court) in the "Circumstances" section of the judgments, but pleadings themselves were not included as inputs.[29] The law section, which "considers the merits of the case, through the use of legal argument," was also input into the model to determine how well that section alone could "predict" the case outcome.[30]

Aletras et al. were selective in the corpus they fed to their algorithms. The only judgments that were included in the corpus were those that passed both a "prejudicial stage" and a second review.[31] In both stages, applications are denied if they do not meet "admissibility criteria," which are largely procedural in nature.[32] To the extent such procedural barriers are deemed "legal," we might immediately identify a bias problem in the corpus—that is, the types of cases where the law entirely determined the outcome (no matter how compelling the facts may have been) were removed from a data set ostensibly fairly representative of the universe of cases

generally. This is not a small problem, either: the overwhelming majority of applications are deemed inadmissible or struck out, and are not reportable.[33]

But let us assume, for now, that the model only aspires to offer data about the realist/formalist divide in cases that do meet admissibility criteria. There are other biases in the data set. Only cases that were in English, approximately 33% of total ECHR decisions, are included.[34] This is a strange omission, since the NLP approach employed here has no semantic content—that is, the *meaning* of the words does not matter to it. Presumably this omission arose out of concerns to make data coding and processing easier.

There is also a subject matter restriction that further limits the scope of the sample. Only cases addressing issues in Articles 3, 6, and 8 of the Convention were included in training and in verifying the model. And there is yet another limitation: the researchers then threw cases out randomly (so that the dataset contained an equal number of violation/no violation cases) before using them as training data.[35]

III. Problematic Characteristics of the ECHR Textual "Predictive" Model

The algorithm used in the case depended on an atomization of case language into words grouped together in sets of one, two, three, and four-word groupings, called N-grams.[36] Then, the most frequent 2,000 N-grams, not taking into consideration "grammar, syntax and word order," were placed in feature matrices for each section of decisions and for the entire case by using the vectors from each decision.[37] Topics, which are created by "clustering together n-grams", were also created.[38] Both topics and N-grams were used to "to train Support Vector Machine (SVM) classifiers." As the authors explain, an "SVM is a machine learning algorithm that has shown particularly good results in text classification, especially using small data sets."[39]

Model training data from these opinions were "n-gram features," which consist of groups of words that "appear in similar contexts."[40] Matrix mathematics, which are manipulations on 2-dimensional tables, and vector space models, which are based on a single column within a table, were programmed to determine clusters of words that should be similar to one another based on textual context.[41] These clusters of words are called topics. The model prevented a word group from showing up in more than one topic. Thirty topics, or sets of similar word groupings, were also created for entire court opinions. Topics were similarly created for entire opinions for each

article.[42] As the Court opinions all follow a standard format, the opinions could be easily dissected into different identifiable sections.[43]

Note that these sorting methods are as legally meaningless as they are mathematically sophisticated.[44] N-grams and topics are not sorted the way a treatise-write might try to organize cases, or a judge might try to parse divergent lines of precedent. Rather, they simply serve as potential independent variables to predict a dependent variable (was there a violation, or was there not a violation, of the Convention).

Before going further into the technical details of the study, it is useful to compare it to prior successes of machine learning, in facial or number recognition. When a facial recognition program successfully identifies a given picture as an image of a given person, it does not necessarily achieve that machine vision in the way a human being's eye and brain would do so. Rather, an initial, training set of images (or perhaps even a single image) of the person are processed, perhaps on a 1,000 by 1,000-pixel grid. Each box in the grid can be identified as either skin or not-skin, smooth or not-smooth, along hundreds or even thousands of binaries—many of which would never be noticeable by the human eye. Moreover, such parameters can be related to one another—so, for example, regions hued as "lips" or "eyes" might have a certain maximum length or width—or, ratio to one another (such that a person's facial "signature" reliably has eyes that are 1.35 times as long as they are wide). Add up enough of these ratios for easily recognized features (ears, eyebrows, foreheads, etc.), and software can quickly find a set of mathematical parameters unique to a given person—or at least unique enough that an algorithm can predict a given picture is, or is not, a picture of a given person, with a high degree of accuracy. The technology found early commercial success with banks, which needed a way to recognize numbers on checks (given the wide variety of human handwriting). With enough examples of written numbers (properly reduced to data via dark or filled spaces on a grid), and computational power, that recognition can become near-perfect.

Before assenting too quickly to the application of such methods to words in cases (as we see them applied to features of faces), we should note that there are not professions of "face recognizers," or "number recognizers," among human beings. So while Facebook's face recognition algorithm, or TD Bank's check sorter, do not obviously challenge our intuitions about how we recognize faces or numbers, applying ML to legal cases should be marked as a

jarring imperialism of ML methods into domains associated with a rich history of meaning (and, to use a classic term from the philosophy of social sciences, *Verstehen*). In the realm of face recognizing, "whatever works" as a pragmatic ethic of effectiveness underwrites some societies' acceptance of width/length ratios and other methods to assure algorithmic recognition of individuals.[45] The application of ML approaches devoid of apprehension of meaning in the legal context is more troubling.

For example, Aletras et al. acknowledge that there are cases where the model predicts the incorrect outcome because of the similarity in words in cases that had opposite results. In that case, even if information regarding specific words that triggered the SVM classifier were output, users might not be able to easily determine that the case was likely misclassified.[46] Even with confidence interval outputs, this type of problem does not appear to have an easy solution. This is particularly troubling for due process if such an algorithm, in error, incorrectly classified someone's case because it contained language similarities to another very different case.[47] When the cases are obviously misclassified in this way, models like this would likely "surreptitiously embed biases, mistakes and discrimination, and worse yet, even reiterate and reinforce them on the new cases processed."[48] So, too, might a batch of training data representing a certain time period when a certain class of cases was dominant, help ensure the dominance of such cases in the future. For example, the "most predictive topic" for Article 8 decisions included prominently the words "son, body, result, russian." If the system were used in the future to triage cases, *ceteris paribus*, it might prioritize cases involving sons over daughters, or Russians over Poles.[49] But if those future cases do not share the characteristics of the cases in the training set that led to the "predictiveness" of "son" status or "Russian" status, their prioritization would be a clear legal mistake. Indeed, they could result in a denial of due process, which guarantees each individual a right for their case to be heard on its own merits, rather than instantly assimilated (on inarticulable grounds) into some broader category.

Troublingly, the entire "predictive" project here may be riddled with spurious or biased correlations.[50] As any student of statistics knows, if one tests enough data sets against one another, spurious correlations will emerge. For example, Tyler Vigen has shown a very tight correlation between the divorce rate in Maine and per capita consumption of margarine between 2000 and 2009.[51] It is unlikely that one variable there is driving the other. Nor is it likely that some intervening variable is affecting both butter consumption and divorce rates in a similar

way, to ensure a similar correlation in the future. Rather, this is just the type of random association one might expect to emerge once one has thrown enough computing power at enough data sets.

It is hard not to draw similar conclusions with respect to Aletras et al.'s "predictive" project. Draw enough variations from the "bag of words," and some relationships will emerge. Given that the algorithm only had to predict "violation" or "no-violation," even a random guessing program would be expected to have a 50% accuracy rate. A thought experiment easily deflates the meaning of their trumpeted 79% "accuracy." Imagine that the authors had continual real time surveillance of every aspect of the judges' lives before they wrote their opinions: the size of the buttons on their shirts and blouses, calories consumed at breakfast, average speed of commute, height and weight, and so forth. Given a near infinite number of parameters of evaluation, it is altogether possible that they could find that a cluster of data around breakfast type, or button size, or some similarly irrelevant characteristics, also added an increment of roughly 29% accuracy to the baseline 50% accuracy achieved via randomness (or always guessing violation). Should scholars celebrate the "artificial intelligence" behind such a finding? No. Ideally they would chuckle at it, as readers of Vigen's website find amusement at random relationships between, say, number of letters in winning words at the National Spelling Bee, and number of people killed by venomous spiders (which enjoys a 80.57% correlation).

This may seem unfair to Aletras et al., since they are using so much more advanced math than Vigen is. However, their models do not factor in meaning, which is of paramount importance in rights determinations. To be sure, words like "burial," "attack," and "died" do appear properly predictive, to some extent, in Article 8 decisions, and cause no surprise when they are predictive of violations.[52] But what are we to make of inclusion of words like "result" in the same list? There is little to no reasoned explanation in their work, as to why such words *should* be predictive with respect to the corpus, let alone future case law.

This is deeply troubling, because it is a foundational principle of both administrative and evidence law that irrelevant factors should not factor into a decision. To be sure, there is little reason the ECHR would use such a crude model to determining the outcome of cases before it, or even to use it as a decision aide. However, software applications often are used in ways that they were not intended. When they are billed as predictive models, attorneys and others could likely

use the models for their own triage purposes. This is especially dangerous when attorneys are generally not very familiar with statistical analysis and machine learning. The legal community's ability to scrutinize such models, and correctly interpret their results, is questionable.[53] Journalistic hype around studies like this one shows that public understanding is likely even more impaired.[54]

Aletras et al. are aware of many problems with their approach, and continually hedge about its utility in the paper. But they still assert:

> Overall, we believe that building a text-based predictive system of judicial decisions can offer lawyers and judges a useful assisting tool. The system may be used to rapidly identify cases and extract patterns that correlate with certain outcomes. It can also be used to develop prior indicators for diagnosing potential violations of specific Articles in lodged applications and eventually prioritise the decision process on cases where violation seems very likely. This may improve the significant delay imposed by the Court and encourage more applications by individuals who may have been discouraged by the expected time delays.[55]

The paper's abstract claims the model "can be useful, for both lawyers and judges, as an assisting tool to rapidly identify cases and extract patterns which lead to certain decisions." Aletras, in a podcast interview, also stated that the model could be used for case triage.[56] However, a judicial system that did so, without attending to all the critiques we have developed above (and perhaps many more), would seriously jeopardize its legitimacy.

For example, consider how non-representative the training data here is. Aletras et al. openly acknowledges a potential issue with "selection effect", or the ability of the model to be useful to the multitude of cases that were dismissed before being heard by the Grand Chamber.[57] Petitions that were determined to be inadmissible before trial were not included in this study, as they are "not reported." Therefore, the model's output is narrowed significantly. Despite these problems, there is a danger that the model could be deployed by bureaucrats at the ECHR to prioritize certain petitions, given that the Court is deluged with thousands of petitions each year, and can only decide a fraction of those cases. Without a clear understanding of how the model is predicting the success of a claim, it would be irresponsible for judges or their clerks or subordinates to use it.[58]

IV. Conclusion

This article has explored shortcomings of Aletras et al.'s *Predicting Judicial Decisions*, to question the nature, purpose, and likely results of computational analysis of legal corpora. We believe the social value of pure ML/NLP-based prediction of judicial decisions, divorced from analysis of the meaning of the texts analysed, is frequently overstated. We also believe that such predictive tools are, at present, largely irrelevant to debates in jurisprudence. If they continue to gloss over the question of social and human meaning in legal systems, NLP researchers should expect justified neglect of their work by governments, law firms, businesses, and the legal academy.[59]

Efficiency *simpliciter* is not an adequate rationale for modeling predictions of judicial behavior. Nor are such approaches' potential to generate more complaints and litigation (where success is predicted) or to discourage such interventions (where success is not predicted) necessarily a positive affordance for society as a whole. Critics of machine learning have long complained that the bias in corpora of past data may simply be recycled into bias in future predictions. Heretofore, authors who have attempted to model the future behavior of courts from their past behavior have given us little sense of how well such biases may be counteracted, or even detected, once their approaches are more widely disseminated.

The legal trade press often heralds computational modeling of judicial behavior as an advance in access to justice: the better we (or, more precisely, those with access to the right data sets and other resources) can predict judgments, so the thinking goes, the better we can know what penalties law-breaking will result in. But technophiles underestimate the degree to which inequality in access to human attorneys is exacerbated by current inequality in access to the best software, legal analytics, and other technology, and by the many ways in which past inequality deeply shaped training data.

To be sure, predictive analytics in law may improve over time. But that possibility does not undermine our position. The most important contribution of our critique is not to cast doubt on the likelihood of further advances in algorithms' predictive power; rather, we question where the results of such projects are useful to the legal system, and demonstrate how they threaten to undermine its legitimacy. The pragmatic and the critical uses of predictive algorithms are deeply in tension. An analyst may reveal biases in judgments, such as legally irrelevant details that

somehow seem to be correlated with, and perhaps even driving, decisions. The same analyst may sell the predictive tool to attorneys or courts, as a case selection or triage tool. But precisely to the extent past training data reflect bias, they are likely to reinforce and spread the influence of that bias when they are utilized by actors outside the judicial system (who may, for example, not even try to advocate for a particular class of meritorious cases, since decisionmakers are systematically biased against them). Academics should never assume that merely increasing the ability to predict the future (or analyze what was most important in decisions of the past) is an unalloyed good. Rather, a long history of social scientific research on reflexivity reveals how easily such analysis can exacerbate, rather than resolve, the problems it reveals.

To the extent such reflexivity develops, better that the Pandora's Box of legal predictive analytics had never been opened. Machine learning may simply replay regrettable aspects of the past into the future. On the other hand, once robust predictive models are available, jurisdictions should carefully consider rules to level the playing field, to ensure that all parties to a dispute have access to critical technology. The law itself is free to consult and copy. To the extent that legal technology determines or heavily influences advocacy, it, too, should be open on equal terms to all parties to a dispute. And at the very least, any deployment of such approaches during litigation should be revealed to the judge presiding over it, and to opposing parties, when it is deployed. Such a general rule of disclosure is vital to future efforts to understand the influence of machine learning, artificial intelligence, and predictive analytics, on the legal system as a whole.

Notes

[1] We wish to thank Julia Powles, Andrew Selbst, and Will Bateman for commenting on an earlier draft.
[2] Mireille Hildebrandt, 'Law as Computation in the Era of Artificial Legal Intelligence: Speaking Law to the Power of Statistics' (2017) UTLJ <page number to come> [Hildebrandt, 'Law as Computation'].
[3] Frank Pasquale, *The Black Box Society* (Cambridge, MA: Harvard University Press, 2015); Ariel Ezrachi & Maurice Stucke, *Virtual Competition* (Cambridge, MA: Harvard University Press, 2016); Hildebrandt, 'Law as Computation,' supra note 2 at 11 ("We are now living with creatures of our own making that can anticipate our behaviours and pre-empt our intent. They inform our actions even if we don't know it and their inner workings seem as opaque as our own unconscious.").
[4] Jason Tashea, 'Risk-Assessment Algorithms Challenged in Bail, Sentencing and Parole Decisions,' ABA J (1 Marc 2017), online: <http://www.abajournal.com/magazine/article/algorithm_bail_sentencing_parole>.
[5] Chad Oldfather, 'Writing, Cognition, and the Nature of the Judicial Function' (2007) 96 Geo LJ 1283 (discussing the types of decisions that must be justified in writing); see also Simon Stern, 'Copyright Originality and Judicial Originality' (2013) 63 UTLJ 385 (discussing the ways in which judicial writing can achieve its justificatory function).
[6] BF Skinner, *Beyond Freedom and Dignity* (Hardmondsworth, UK: Penguin Books Ltd, 1971).
[7] For a powerful defense of persuasion and other forms of rhetoric common in legal and political contexts, see Bryan Garsten, *Saving Persuasion* (Cambridge: Harvard University Press, 2006).
[8] Brian Sheppard, 'Why Digitizing Harvard's Law Library May Not Improve Access to Justice,' Bloomberg BigLaw Business (12 November 2015), online: <https://bol.bna.com/why-digitizing-harvards-law-library-may-not-improve-access-to-justice/> ("Ravel has already said that the users will not have free access to its most powerful analytic and research tools. Those will exist behind a paywall."). Ravel Law has been described as an "artificial intelligence company" that "provides software which can predict the arguments which would win over a judge." 'RavelLaw Acquired by LexisNexis,' Global Legal Post (9 June 2017), online: http://www.globallegalpost.com/big-stories/ravel-law-acquired-by-lexisnexis-32302305/; J Dixon, 'Review of Legal Analytics Platform,' LitigationWorld (23 September 2016), online: <https://lexmachina.com/wp-content/uploads/2016/10/LitigationWorld-Review-2016.pdf> (describing cost of predictive analytics platform for patent cases); Frank Pasquale, 'Technology, Competition, and Values' (2007) 8 Minn JL Sci & Tech 607, 608 (describing inequality-enhancing effects of technological arms races).
[9] Oliver Wendell Holmes, 'The Path of the Law' (1897) 110 Harv L Rev 991; Ian Kerr, 'Chapter 4: Prediction, Preemption, Presumption: The Path of Law After the Computational Turn,' in *Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology* (Abingdon, Oxon, [England]; New York: Routledge, Taylor & Francis, 2013), at 91–120.
[10] Hildebrandt, 'Law as Computation,' supra note 2 at [page].
[11] Ibid.
[12] Blaise Agüera y Arcas, Margaret Mitchell and Alexander Todorov, 'Physiognomy's New Clothes,' Medium (6 May 2017), online: <https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a>; Danah Boyd & Kate Crawford, 'Critical Questions for Big Data' (2012) 15:5 Info Comm & Society 662.
[13] Federico Cabitza, 'The Unintended Consequences of Chasing Electric Zebras,' IEEE SMC Interdisciplinary Workshop HUML 2016, The Human Use of Machine Learning (16 December 2016, Venice, Italy), online: <https://www.researchgate.net/publication/311702431_The_Unintended_Consequences_of_Chasing_Electric_Zebras> [Cabitza, 'Unintended Consequences'] ("ML approach risk[s] to freeze into the decision model two serious and often neglected biases: selection bias, occurring when training data (the above experience E) are not fully representative of the natural case variety due to sampling and sample size; and classification bias, occurring when the single categories associated by the raters to the training data

oversimplify borderline cases (i.e., cases for which the observers do not agree, or could not reach an agreement), or when the raters misclassify the cases (for any reason). In both cases, the decision model would surreptitiously embed biases, mistakes and discriminations and, worse yet, even reiterate and reinforce them on the new cases processed.”).

[14] Jathan Sadowski & Frank Pasquale, ‘The Spectrum of Control: A Social Theory of the Smart City,’ First Monday (31 Aug 2015), online: <http://firstmonday.org/article/view/5903/4660>

[15] Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preotiuc-Pietro, & Vasileios Lampos, ‘Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective’ (2006) 2 PeerJ Comp Sci 92. [Aletras et al, ‘Predicting Judicial Decisions’]

[16] Ibid.

[17] Aletras is quoted in Anthony Cuthbertson, ‘Ethical Artificial Intelligence “Judge” Predicts Human Rights Trials,’ Newsweek (24 October 2016), online: <http://www.newsweek.com/ethical-artificial-intelligence-judge-predicts-human-rights-trials-513012>. It seems clear from context that Aletras includes this study in his hopes for AI, as it does in his interview for the Australian podcast *Future Tense*, infra note 53.

[18] Aletras et al, ‘Predicting Judicial Decisions,’ supra note 15 at [page] (“we highlight ways in which automatically predicting the outcomes of ECtHR cases could potentially provide insights on whether judges follow a so-called *legal model* of decision making or their behavior conforms to the *legal realists' theorization*, according to which judges primarily decide cases by responding to the stimulus of the facts of the case.”)

[19] As Yoshua Bengio explains, “Another big challenge is natural language understanding. . . . it’s still not at the level where we would say the machine understands. That would be when we could read a paragraph and then ask any question about it, and the machine would basically answer in a reasonable way, as a human would. We are still far from that.” Will Knight Interview with Yoshua Bengio, “Will Machines Replace Us,” *MIT Technology Review*, Jan. 29. 2016.

[20] See European Court of Human Rights, *Questions & Answers,* (Council of Europe) at 1, 4, online: <http://www.echr.coe.int/Documents/Questions_Answers_ENG.pdf>; Aletras et al, ‘Predicting Judicial Decisions,’ *supra* note 15.

[21] Aletras et al, ‘Predicting Judicial Decisions,’ *supra* note 15 (“Our models can reliably predict ECtHR decisions with high accuracy, i.e., 79% on average.”).

[22] ‘Could AI Replace Judges and Lawyers?’ BBC News (24 October 2016), online: <http://www.bbc.com/news/av/technology-37749697/could-ai-replace-judges-and-lawyers>

[23] Aletras et al, ‘Predicting Judicial Decisions,’ *supra* note 15, at 4. Data from this section, or a combination of it with “Topics,” generated the predictive accuracy that was trumpeted in the paper itself, and numerous media accounts of it as a stepping stone to substitutive automation. We focus in this section only on the best-performing aspects of the model; our critiques apply *a fortiori* to worse-performing ones.

[24] To expand the analogy: the analysis of food on a molecular level, disconnected from taste, appearance, or other secondary qualities perceptible by humans, is parallel to NLP’s “bag of words” approach to processing a text on the level of individual words disconnected from meaning. We use this classic reference to hard-core legal realism’s gustatory characterization of irrational indeterminacy in adjudication, in honor of Aletras et al.’s claims to buttress the legal realist approach. Jerome Frank, *Courts on Trial* (Princeton: Princeton University Press, 1973), at 162 (“Out of my own experience as a trial lawyer, I can testify that a trial judge, because of overeating at lunch, may be somnolent in the afternoon court-session that he fails to hear an important item of testimony and so disregards it when deciding the case.”).

[25] Ibid at 3.

[26] Ibid at 4, emphasis added. The authors themselves acknowledge that “The choices made by the Court when it comes to formulations of the facts incorporate implicit or explicit judgments to the effect that some facts are more relevant than others. This leaves open the possibility that the formulations used by the Court may be tailor-made to fit a specific preferred outcome.” Ibid. They then give some reasons to

believe that this stacking-the-deck effect is "mitigated," but give no clear sense of how to determine the degree to which it is mitigated.

[27] Indeed, some legal realists may assume that it is very difficult to identify the facts driving judges' decisions in their written opinions, since they believe judges are skilled at obscuring both their ideology and fact-driven concerns with complex legal doctrine. One need not adopt a Straussianly esoteric hermeneutics to understand the challenges such a view poses to the very text-based analytics deemed supportive of it by Aletras et al.

[28] Aletras et al. conclude that their study supports the proposition that the Court's decisions are "significantly affected by the stimulus of the facts." Ibid. This "finding," such as it is, is a rather trivial one without further explanation. Judicial proceedings not "significantly affected by" facts would be arbitrary and lawless. The term "stimulus" evokes behaviorist logic of mind as machine. What matters is not so much the reasoning in the cases, but the "facts" considered as bare stimulus, an input processed by the judicial system into an output of decision.

[29] See Aletras et al, 'Predicting Judicial Decisions,' supra note 15.

[30] See ibid.

[31] See ibid.

[32] See ibid.

[33] See ibid. Moreover, focus on an appellate court like the ECHR also biases the outcome in favor of realism, since clear opportunities for the application of law are almost certainly resolved at lower levels of the judicial system.

[34] European Ct of Human Rights, 'HUDOC' (10 Sept. 2017) online: <https://hudoc.echr.coe.int/eng#{"documentcollectionid2":["GRANDCHAMBER","CHAMBER"]}> (18,438 of 55,571 case decisions as of Sept. 10, 2017 were in English).

[35] See Aletras et al, 'Predicting Judicial Decisions,' supra note 15; Nikolaos Aletras et al, 'ECHR Dataset,' (Figshare), online: <https://figshare.com/s/6f7d9e7c375ff0822564>; Christopher D. Manning, Prabhakar Raghavan & Himrich Schütze, *Introduction to Information Retrieval* (Cambridge University Press, 2008) online: <https://nlp.stanford.edu/IR-book/html/htmledition/large-and-difficult-category-taxonomies-1.html>; Andrea Dal Pozzolo et al, 'Calibrating Probability with Undersampling for Unbalanced Classification', 2015 IEEE Symposium Series on Computational Intelligence, online: <https://www3.nd.edu/~rjohns15/content/papers/ssci2015_calibrating.pdf > (Undersampling to achieve an equal number of violation/no-violation like what is done here is typical with "unbalanced datasets" in SVM, the algorithm used in Aletras et al).

[36] See Aletras et al, 'Predicting Judicial Decisions,' supra note 15; Kavita Ganesan, 'What are N-Grams?' Text Mining, Analytics, & More (23 November 2014), online: <http://text-analytics101.rxnlp.com/2014/11/what-are-n-grams.html>. [Ganesan, 'What are N-Grams?']

[37] See Aletras et al, 'Predicting Judicial Decisions,' supra note 15.

[38] See ibid.

[39] See ibid.

[40] See Aletras et al, 'Predicting Judicial Decisions,' supra note 15; See Ganesan, 'What Are N-Grams?' supra note 34 ("N-grams of texts are extensively used in text mining and natural language processing tasks. They are basically a set of co-occurring words within a given window and when computing the n-grams you typically move one word forward [although you can move X words forward in more advanced scenarios]. For example, for the sentence 'The cow jumps over the moon'. If N=2 (known as bigrams), then the ngrams would be: the cow; cow jumps; jumps over; over the; the moon. So you have 5 n-grams in this case. Notice that we moved from the->cow to cow->jumps to jumps->over, etc, essentially moving one word forward to generate the next bigram. If N=3, the n-grams would be: the cow jumps; cow jumps over; jumps over the; over the moon. So you have 4 n-grams in this case.").

[41] See Aletras et al, 'Predicting Judicial Decisions,' supra note 15.

[42] See ibid.

[43] See ibid.

[44] Legal meaninglessness here denotes an absence of appropriate explainability, as to why even a refined algorithm decided cases in the way it did. For more on this problem of explainability, *see* Frank Pasquale, 'Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society,' 78 Ohio State Law Journal (forthcoming, 2017), online: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3002546>.
[45] We should note that this acceptance is not universal—wary of violating some jurisdictions laws, Facebook does not operate the face recognition algorithm automatically in them.
[46] See Aletras et al, 'Predicting Judicial Decisions,' supra note 15 ("On the other hand, cases have been misclassified mainly because their textual information is similar to cases in the opposite class.").
[47] Cabitza, 'Unintended Consequences," supra note 13; Hildebrandt, 'Law as Computation' supra note 2 at [11] ("I will discuss four implications that may disrupt the concept [of] the Rule of Law: (1) the opacity of ML software may render decisions based on its output inscrutable and thereby incontestable; (2) the shift from meaningful information to computation entails a shift from reason to statistics, and from argumentation to simulation; (3) in the process of developing and testing data driven legal intelligence a set of fundamental rights may be infringed, compromised or even violated, notably the right to privacy, to non-discrimination, to the presumption of innocence and due process, while also impacting consumer and employee protection and competition law. Finally, I argue that, (4) to the extent that the algorithms become highly proficient – due to being trained by excellent domain experts in law – lawyers may outsource part of their work, as a result of which they may deskill as the software achieves high levels of accuracy.").
[48] Cabitza, 'Unintended Consequences,' supra note 13.
[49] Aletras et al, 'Predicting Judicial Decisions,' supra note 15 at 16.
[50] On the question of bias, see Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan, 'Semantics derived automatically from language corpora contain human-like biases,' Science 356, 183-186 (14 April 2017); see generally Ethics in Natural Language Processing: Proceedings of the First ACL Workshop, EACL 2017, online: <http://www.aclweb.org/anthology/W17-16>; Tania Cerquitelli, Daniele Quercia, and Frank Pasquale, eds., Transparent Data Mining for Big and Small Data (2017).
[51] Tyler Vigen, 'Spurious Correlations,' online: <http://www.tylervigen.com/spurious-correlations>. Vigen, then "a criminology student at Harvard Law School, wrote a computer programme to mine datasets for statistical correlations. He posts the funniest ones to Spurious Correlations," a website. James Fletcher, Spurious Correlations: Margarine Linked to Divorce?' BBC News (26 May 2014), online: http://www.bbc.com/news/magazine-27537142.
[52] Aletras et al, 'Predicting Judicial Decision,' supra note 15 at 15.
[53] Hildebrandt, 'Law as Computation,' supra note 2 at 12 ("Whereas most of us have learnt to read and write, we were not trained to 'read' and 'write' statistics, we cannot argue against the assumptions that inform ML applications, and we miss the vocabulary that frames 'training sets', 'hypotheses space', 'target functions', 'optimization', 'overfitting' and the more. So, even if experts manage to verify the software, most of us lack the skills to make sense of it; we may in fact be forced to depend on claims made by those who stand to gain from its adoption and on the reputation of those offering this type of data driven legal services. This is also the case when we buy and drive a car, but the reliability here is easier to test. We recognize a car crash and would not appreciate a car that drives us from A to C if we want to get to B; with legal intelligence we may simply not detect incorrect interpretations.").
[54] Michael Byrne, 'How to Navigate the AI Hypestorm,' The Outline (15 September 2017), online: <https://theoutline.com/post/2248/how-to-navigate-the-coming-a-i-hypestorm> ("[T]he basic playbook is to take a bunch of examples of some phenomenon to be detected, reduce them to data, and then train a statistical model. Faces [and cases] reduce to data just like any other sort of image [or text] reduces to data. . . . [If the] machine learning model [is] making better predictions than humans, that's . . . completely meaningless. Like, obviously the computer is going to do a better job because it's solving a math problem and people are solving a people problem.").
[55] Aletras et al, 'Predicting Judicial Decisions,' supra note 15 at 3.

[56] This use was suggested in a podcast interview. Future Tense, 'Augmented eternity and the potential of prediction' (12 March 2017), online: http://www.abc.net.au/radionational/programs/futuretense/augmented-eternity-and-the-potential-ofprediction/8319648 ("Nikolaos Aletras: We can use textual evidence to predict the cases tried in major international courts, and also we set some lights on what drives judicial decisions. And we think that we can provide then a system to lawyers and judges to prioritise cases easily and in a fast way."). While correctly insisting that the model could not be used to replace judges, its potential to prioritize cases for consideration was discussed.

[57] See Aletras et al, 'Predicting Judicial Decisions,' supra note 15 (the "selection effect" that "pertains to cases judged by the ECtHR as an international court. Given that the largest percentage of applications never reaches the Chamber or, still less, the Grand Chamber, and that cases have already been tried at the national level, it could very well be the case that the set of ECtHR decisions on the merits primarily refers to cases in which the class of legal reasons, defined in a formal sense, is already considered as indeterminate by competent interpreters. This could help explain why judges primarily react to the facts of the case, rather than to legal arguments. Thus, further text-based analysis is needed in order to determine whether the results could generalise to other courts, especially to domestic courts deciding ECHR claims that are placed lower within the domestic judicial hierarchy.").

[58] Moreover, difference in legal knowledge and competence between attorneys and pro se litigants is often an important factor in case outcomes. What if differences in resources drove some of the differences in outcomes in the training data here? The types of claims prioritized by expensive attorneys (or those able to afford them) could end up as part of a template for potential future winning (or expedited) claims, further stratifying litigants. For more on the importance of resources, see Judicial Council of California, *Handling Cases Involving Self-Represented Litigants* (San Francisco: Judicial Council of California, 2017), online: <http://www.courts.ca.gov/documents/benchguide_self_rep_litigants.pdf> ("self-represented litigants often have difficulty preparing complete pleadings, meeting procedural requirements, and articulating their cases clearly to the judicial officer. These difficulties produce obvious challenges.").

[59] Premature quantification, metricization, and algorithmatization all share this allergy to interpretation and meaning. See, e.g., Christopher Newfield and Heather Steffen, 'Remaking the University: Metrics Noir,' *Los Angeles Review of Books*, Oct. 11, 2017 (commenting on "a particularly subtle and difficult limit of the numerical: its aversion to the interpretive processes through which the complexities of everyday experiences are assessed. Physical and mental states, injuries and attitudes toward them, people in variable social positions always appear together, and their qualities need to be sorted out."); Frank Pasquale, 'Professional Judgment in an Era of Artificial Intelligence and Machine Learning,' *Boundary2* (forthcoming, 2018); Mark D. White, *The Decline of the Individual* (Palgrave, 2017) (discussing Sorokin's critique of "quantophrenia").