2019

The Impact of Artificial Intelligence on Rules, Standards, and Judicial Discretion

Frank Fagan

Saul Levmore

THE IMPACT OF ARTIFICIAL INTELLIGENCE ON RULES, STANDARDS, AND JUDICIAL DISCRETION

Frank Fagan & Saul Levmore[*]

93 SOUTHERN CALIFORNIA LAW REVIEW __ (forthcoming 2019)

Artificial intelligence (AI), and machine learning in particular, promises lawmakers greater specificity and fewer errors. Algorithmic lawmaking and judging will leverage models built from large stores of data that permit the creation and application of finely tuned rules. AI is therefore regarded as something that will bring about a movement from standards to rules. Drawing on contemporary data science, this Article shows that machine learning is less impressive when the past is unlike the future, as it is whenever new variables appear over time. In the absence of regularities, machine learning loses its advantage and, as a result, looser standards can become superior to rules. We apply this insight to bail and sentencing decisions, as well as familiar corporate and contract law rules. More generally, we show that a Human-AI combination can be superior to AI acting alone. Just as today's judges overrule errors and outmoded precedent, tommorrow's lawmakers will sensibly overrule AI in legal domains where the challenges of measurement are present. When measurement is straightforward and prediction is accurate, rules will prevail. When empirical limitations such as overfit, Simpson's Paradox, and omitted variables make measurement difficult, AI should be trusted less and law should give way to standards.

I. INTRODUCTION

In earlier generations, office workers and homemakers became more efficient by learning to work with machines. In the present era, our adjustment to human-machine partnerships is more challenging because today's machines, unlike early typewriters, word-processing software, and washing machines, can now make decisions and improve on their own. Humans will soon need to decide on a regular basis when to trust and when to overrule machines. In the case of law, as in the case of autonomous

[*] Fagan is Associate Professor of Law, EDHEC Business School, France; Levmore is the William B. Graham Distinguished Service Professor of Law at the University of Chicago Law School. We are grateful for the thoughtful comments we received from William Hubbard, Michael Livermore, and Christophe Croux, as well as participants of the University of Chicago School of Law faculty workshop.

vehicles and other endeavors, part of this difficulty is that the most promising advances in AI (artificial intelligence) are opaque; errors are harder to identify because faulty assumptions and omitted variables are difficult to identify inasmuch as algorithms cannot be counted on to explain how they reached their decisions. This Article explores the division of labor between AI and human lawmakers. First, we suggest that in stable legal environments, humans should be more hesitant to second-guess machines than should machines be discounted when they differ with judges and other human lawmakers. More generally, we show that AI, and sensible overruling in both directions, will lead to a different mix of rules and standards in law. In addition, we explain why AI, working alone, has its flaws, and we do so with examples that touch on pre-trial bail, parole, and sentencing decisions, and that draw on the laws of torts, contracts, bankruptcy, and corporations. We show why machines will turn some standards into rules, and will add specificity to familiar legal rules. In exceptional but interesting cases, an understanding of the limits of machine learning will have the reverse effect, as it will encourage law to substitute some standards—and human judgment—in place of familiar and precise legal rules.

II. THE COMBINATORY POWER OF HUMANS AND MACHINES

A. The Advent of Combinations

Combinations of, or teamwork between, people and their machines is not new. Forty years ago it would have been obvious that even the most talented secretary needed to have good typing skills. The secretary was judged by an ability to screen office visitors and to take dictation and improve correspondence, among other things, but these contributions were especially valued if the employee could type accurately and quickly. There were advantages of combining humans into teams of workers, but also of combining the efforts of an employee and a typewriter—a human and a machine. The two were complements more than substitutes. As the machine improved, the relative contribution of the human did not necessarily decrease, and indeed compensation increased until computers enabled more senior people to rely very little on secretaries.

There is reason to think that the same is true in the modern era with respect to the relationship between humans and computers, although this combination is more interesting because some tasks and decisions can be made by the machine alone, while even the IBM Selectric typewriter required a human to do anything at all. Machine learning (ML), a subset of

artificial intelligence, is a term applied to the ability of modern machines to improve on their own, after some programming about the goals sought by humans and the sources of data offered to machines.[1] The IBM typewriter was mechanical; a modern word processing program is a simple example of AI, and it uses structured data[2] to correct misspelled words. Moreover, the word processor engages in machine learning if it is able to suggest sentences based on the users' previous e-mails to similar parties. The more data, in the form of previous messages, user corrections, or favorable responses by recipients, the better the suggestions are likely to be.

Humans were impressed when machine learning enabled a machine to

[1] Artificial intelligence (AI) is a general term that includes things that machines have done for years. In this Article we use it in a way that refers to the transfer of decisionmaking, or investigation of facts, away from humans. Machine learning (ML) is best understood as a subset of AI, as it draws attention to the capacity of a program to improve on its own. In its extreme, unstructured machine learning, humans initiate the AI's work by providing goals as well as data or some instructions about where and how to look for data, but the machine is free to make connections, reach conclusions, or look for more data in ways that the human had not contemplated. In many cases, AI and ML are expressions that can be used interchangeably. But even this description has been misleading because there are many approaches within the world of machine learning. An algorithm that imitates evolution, for instance, is not exactly looking for previously unidentified connections. Moreover, as explained and incorporated below, a machine that looks only for connections is apt to run into problems of overfit. But inasmuch as surprising connections have captured the attention of legal scholars, we use terms consistent with this interest. Indeed, though AI is itself an inclusive term that may not survive the test of time, we use it here because of its current popularity and because it reflects our interest in the division of lawmaking power between humans and machines.

AI learns from feedback, which can be divided into three types. In *supervised learning*, the AI observes some data, such as arrestee characteristics, and a teacher labels the arrestee's bail suitability and amount. Alternatively, an AI might be given video footage and instructed that it has observed a crime. In *unsupervised learning*, AI learns from patterns in the data with no explicit feedback. The learning task primarily consists of identifying clusters. For instance, an AI taxi might gradually learn to distinguish between good and bad traffic days without ever receiving labeled examples from a teacher. *Reinforcement learning* is a combination of the two. An AI might be "punished" by being overruled (and thus identified or "labelled"), but it is left to the AI to examine patterns and deduce why it was overruled. In practice, these distinctions can overlap. In *semi-supervised learning*, an AI may be given partially labeled and unlabeled data and asked to make a prediction. Alternatively, all of the data may be labeled, but in a systematically inaccurate way. *See* STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 694-95 (3d ed. 2010).

[2] When a machine is given structured data, it is offered inputs or labels in order to examine a particular correlation or other hypothesis that has been previously developed. Structured representation of data underlies knowledge-based learning, where there is prior knowledge about the world. *See* RUSSELL & NORVIG, *supra* note 1 at 57-58.

defeat the highest ranked human in Chess and then in the game of Go;[3] they were absolutely dazzled when they learned that the victorious play included moves and strategies that had not occurred to skilled humans even when the latter had ample time and years of experience to investigate the possibilities. It is when the computer does more than capitalize on its ability to follow human instructions quickly that its artificial intelligence is most noteworthy. Some of these performances turn out to be overstated,[4] perhaps because of a desire by humans to signal an appreciation of novelty and progress, but there is reason to think that over time, machine learning will use its experience and growing databases to improve so that humans will cede more decision-making to their machines. Machines will undoubtedly outperform human judges in deciding matters of pre-trial bail, post-conviction parole, and which tax returns can most profitably be audited. It is already the case that only a stubborn adult thinks that he can outperform Waze or Google Maps in finding the fastest route to a destination, or that she can choose the best restaurant by asking a friend rather than consulting an online aggregator—though determining the best online authority for a given individual is at present a question better allocated to humans.

More remarkable is that the combination of a skilled human and a machine is often superior to what even the best equipped machine can do on its own. A machine that defeats humans is often inferior to a machine that a skilled human can override; the combination triumphs. It may not do so indefinitely or in all domains, but it is notable at present, and particularly useful in thinking about legal applications. The success of combinations, or

[3] *See* David Silver et al., *Mastering the Game of Go with Deep Neural Networks and Tree Search*, 529 NATURE 484, 484 (2016).

[4] One important example is the attention paid to research showing that a machine outperformed human judges in predicting crimes or flight during pre-trial periods when an arrested person was released on bail; the gains from artificial intelligence were considerable. *See* John Kleinberg, et. al., *Human Decisions and Machine Predictions*, 133 Q. J. ECON. 237, 237 (2018). Unfortunately, the competition between these human judges deciding pre-trial bail and an algorithm instructed to minimize crimes and flight (measured by a failure to appear at trial) was a bit unfair and proves something quite different from that often attributed to this work. The machine was given evidence about the successes and failures of its earlier decisions, while the judges would only occasionally (and by accident) know whether those they had set free in the past took flight or were again arrested, or convicted. In machine-learning terms, the judges, and the decision-rules that they were applying to the pre-trial bail cases, suffered from lack of supervision and reinforcement and, as a result, performed relatively poorly. In contrast, the algorithm was trained; it was "supervised" and "reinforced" with performance data, and its decision-rules were penalized for failure and rewarded for success. The real message seems to be that more data, and learning about the effect of one's past decisions, is a good idea. This is unsurprising.

teamwork, is easily understood by considering Freestyle Chess, in which humans can consult outside sources within a time limit; the champions, at least for a time, are human-computer partnerships. These combinations defeat the best machines, and of course the best humans, when these entities act alone. Nor is there reason to think that a combination of machines or of humans is superior to one that mixes human with machine. The humans in Freestyle Chess (or centaur play) appear to have the special, learned skill of playing with machines; the chess ranking of these partnered humans is well below that of the leading players in the world, when all are judged on their success as single, unaccompanied players.

This is an important point, and it is worth restating and exploring. Just as the best secretary of yester-year might have been one who learned to be a good typist, so too the most successful freestyle player is apparently one who has learned to interact well with a machine. In the chess context, the victorious human is good at knowing when to overrule the machine. This may not be true in the future, because of the machine's ability to learn, but the Human-AI combination is more likely to remain superior in enterprises like law where the playing field changes over time.[5] It changes over time in chess to the extent that opponents play different moves as they too learn how to react to a machine or freestyle opponent, but in law, war, and the stock market, the field is likely to change more because of external events as well as reactions to moves made by an artificial intelligence or a Human-AI combination. To the extent that AI succeeds because of its ability to learn from the past, this strength is probably less important the more the future will not resemble the past.

B. The Role of Humans

[5] *See infra* note 9. Imagine if the rules of chess changed every few years. In the extreme case, changes to rules would be random and unpredictable, and data would not be useful for prediction. Only if a pattern is discernible will the machine be able to predict winning moves.

In this Article, we set aside the interesting *philosophical question* of how we might know whether the future is like the past. This is, of course, a familiar riddle of induction, often examined by philosophers with questions such as how we can be reasonably sure that the sun will rise tomorrow, posed against the seemingly absurd idea of whether it would have been reasonable for someone in 1796 to think that George Washington would, in the future, always command the U.S. armed forces given that he had always done so in the past. *See* NELSON GOODMAN, FACT, FICTION, AND FORECAST 59 (4th ed. 1983) (discussing this example). But it is also a *practical question*, inasmuch as part of the argument here is that AI's strength requires some similarity between the past (which is to say the available data) and the future, while humans may be better at knowing when to trust AI less based on insights about the likely difference between past and future.

In freestyle tournaments, as well as in medical care and other settings, the human takes charge of the combinatory effort, and can ignore or overrule the machine. The machine can try to overrule the human but ultimately the next move in the freestyle game, the practice of medicine, and the operation of a conventional or autonomous car or airplane, is in the hands of a human. A good human worker knows not to overrule the machine too often, especially when there is no evidence that the machine has erred. The superiority of human-directed combinations is surprising if we think of the machine as powerful simply because of its speedy calculations. This power helps it examine stored data from previous encounters and plan several moves ahead. But the machine's ability to see new routes to success reflects true machine learning. Skilled humans, playing on their own, may be overly inclined to value precedent, and often think that a problem closely resembles one they have seen before. Again, the combination of human and machine may not triumph in the future, so long as the playing field remains sufficiently stable, but it is an important option to consider in lawmaking both because of law's dynamism and, as we will see, because it is apt to be politically attractive compared to unsupervised machines.

1. Goal-Setting and Overrides

At least for the present, we should expect the best lawmakers, including judges, to be individuals who are adept at evaluating the decisions reached by machines. With AI, machines will recommend (rather than finalize or steer clear of) such things as optimal prison sentences, speed limits and investments in their enforcement, visas, and tariffs. One day they might even recommend the winners of political contests, rather than simply predict winners based on humans' previous voting patterns. It is more challenging to list decisions that will be or should be free of AI, than to list areas where machines can be decisive. Machines will be as useful in law as in the private sector. This leaves two important roles for humans. First, humans will presumably decide the goals; just as humans decided what constitutes a victory in chess, so humans will program machines with the weights that artificial intelligence should assign to efficiency, wealth distribution, short and long-term climate change (and, indeed, the applicable discount rate), and so forth.

Second, humans will have the ability to overrule machines in law, as elsewhere. This power is likely to be a political prerequisite for allowing ML into lawmaking in the first place. Over time, combinations of humans and machines will learn how to coordinate and how often to overrule one

another; a human might simply change the goal midstream, in order to alter the machine's conclusion. Humans and machines might compete in their interpretation of results, but progress should be expected. The evolution of autonomous vehicles offers a useful example. Humans have already observed the value of these vehicles and many humans over-emphasize reported accidents involving experimental vehicles, but over time we expect humans to outsource decision-making so long as they retain the power to overrule their machines.[6] Lawmakers who block this progress will seem foolish. Law might help the transition both by holding the owner or manufacturer of the vehicle strictly liable and, more important, by finding defendants liable when they had overruled their machines with disastrous outcomes. On the other hand, various interest groups might cause law to move in the opposite direction, inefficiently requiring the licensing of new machines or requiring every vehicle or every program to pass frequent and costly inspections or even to have two human operators and other safeguards, as is the case for commercial airliners.[7] But competition among jurisdictions can be expected to bring about socially attractive results, including taxation, redistribution, and other considerations of interest to voters and politicians.

2. Data Selection

It is a mistake to think that instructing the AI is always an easy task, and that the human role in this combination of talents is trivial. Consider, for example, the case of pre-trial bail.[8] The law must decide which arrested people to release or detain before trial; incarceration is costly in many ways, but a primary cost of bail is the risk that the individual will engage in wrongdoing that would not have occurred had he or she been incarcerated.[9]

---

[6] While overruling might be barred for individual drivers, a driverless car network operator will retain residual control if not by means of basic design selection.

[7] It has even been suggested that AI should undergo an approvals process administered by a new agency tasked with certifying AI safety. *See* Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J. L. & TECH. 353, 393 (2016).

[8] Pre-trial bail decisions have been a subject of interest in earlier work about AI, see note 4 *supra*; it is also discussed below in Section III.A.1.

[9] A secondary cost is the danger of decreasing the deterrence effect of criminal law. We can ignore it here because we are a long way from using AI to evaluate this effect; inasmuch as each arrestee is unique, machine learning will fail to detect a relationship between a singular feature of the arrestee and aggregate crime levels. Over time, with more observations, detection of this feature and its possible relation to crime becomes more likely. *See* LESLIE VALIANT, PROBABLY APPROXIMATELY CORRECT 61-62 (2013) (asserting that detection and learning cannot occur without an observable regularity or pattern). Note that even if arrestee features are persistent and measurable, learning is not possible if the

In addition, there is the risk of flight. An obvious instruction is to gather sufficient data about persons previously released on bail, and ask the AI to find characteristics that will help predict who should be freed or detained. The human might set a probability threshold, or might simply ask for the same percentage of detentions that human judges allowed, if only to outperform the human judges who acted without AI partners. Either way, the idea is to release (on bail) those who have a low probability of committing crimes or fleeing the jurisdiction.

The AI will examine the characteristics of persons released in the past and discover the factors associated with those who in fact did or did not flee or commit crimes. It might find that pre-trial release proved successful when decided by Judge A more often than it did when the decision was in the hands of Judge B or C. Alternatively, perhaps the amount of bail turns out to be important. Similarly, perhaps some combination of the identity of the judge and the nature of the crime for which there was an arrest turns out to be the best predictor. Finally, perhaps success is associated with a *more* rather than less serious crime—something that is unlikely to occur to humans. An AI is likely to be good at finding such connections if given a reasonably large data set; the work is mostly automated, except to the extent that someone must choose what data to collect and possibly label for the AI to do its work.[10] At this point, it might be useful to use AI to discern Judge A's decisionmaking process, or "intuition"; Judge A might be unable to articulate the process used, and therefore unable to train other judges. Similarly, other judges might observe Judge A in action but be unable to learn how to copy A and achieve a similar level of success. The best judges, like the best athletes and teachers, are often unable to identify the reasons

---

arrestee's environment is constantly changing. *See id.* (asserting that learning cannot occur when the context of a generalization is changing). Valiant helpfully denotes each of these conditions, respectively, as the Learnable Regularity Assumption and the Invariance Assumption. *Id.*

Note that here the question is whether a judge or AI is better at deciding pre-trial releases and bail. There is something of a puzzle in the use of judges at this stage—but then relying on non-judge parole boards later—following conviction and a period of incarceration. The difference might be explained by the fact that judges are given the job when the accused party is in the courtroom anyway, but not otherwise. But the puzzle would be solved in a mysterious fashion if a study showed that a parole board outperformed AI, and it would be disturbing if, to the contrary, parole boards were even worse than judges.

[10] If the AI suggests that Judge A was a superb decisionmaker, a good investigator will suggest data division and further testing, as *some* judge will by chance turn out to make the right decisions. But if Judge A turns out to be successful in the set-aside data as well, our concerns about overfit will be reduced. The problem of overfit is explained and discussed presently in Section III.A.1.

for their successes.

3. Data Availability

Note, however, that even if all this works well, the AI's power has not been fully tapped. The AI was only offered data about the people who were let out on bail. The earlier, human decisionmakers might have kept people in jail who would not have committed crimes if released, but the AI was unable to test their propensity toward crime or flight because they remained in jail. In fact, if human or AI decisionmakers systemically detain individuals with particular untested characteristics (or combinations of characteristics) then the AI cannot learn how that group of people would behave while on bail. This might occur because judges or machines have been instructed by law, or a collective hunch, to always detain people who have committed crimes of type X or Y, or have personal characteristics of type Q or R. Assuming a large enough sample, a statistician would like to randomly select people who were not released, and then let them out on bail. The AI could then discern the characteristics that identify the subset of this group that turned out to be well-behaved.

Other approaches might achieve the same result. Consider that AI is not necessary in the first step; after Judges A, B, and C make their decisions, bail could be randomly granted to a fraction of arrested and retained persons whose progress could be followed. Alternatively, a human could guess other characteristics that judges disfavored, and offer bail to some of them in order to judge their performance. Indeed, if humans make the earlier decisions, we might decide to let AI find the disfavored characteristics, and then select a random sample from this disfavored group to release on bail and to study.[11] There are other variations for lawmakers to contemplate, but the point is that there is a substantial role for humans in thinking about data limitations and goal-setting.

4. Expenditures

[11] The same is true outside of law. Instead of studying successful employees or law students, we (or AI) might study the characteristics of successful employees or lawyers at other firms (or law schools) that ours did not choose, or we might select a random set of rejected applicants to study. This idea is motivated by Matt Levine, *The Robots Learn by Watching Us*, BLOOMBERG (October 11, 2018), https://www.bloomberg.com/view/articles/2018-10-11/the-robots-learn-by-watching-us?srnd=opinion. Note that the hiring and admissions examples are relatively harmless, while the pre-trial bail example runs the risk of extra crimes committed by the control group, which will be blamed on the AI's thirst for data.

To this point, our legal examples have been limited to regulations and criminal law; nothing has been said about expenditures. Consistent with this pattern, it is easy to imagine AI determining or suggesting the speed limit on a bridge. The limit might be different for various kinds of vehicles, including autonomous cars or trucks, and might vary by the time of day, the number of vehicles detected on the bridge, the presence of nearby emergency vehicles, and so forth. It is less likely that AI will determine when and where to build a bridge in the first place. This is not because AI is incapable of taking into account the value of commuters' time, the cost of bridge building, and other variables inserted by humans (with the assistance of AI)—but rather because of interest groups and other political influences. Even in a world where AI instructs to build a bridge of a particular size at a given location, and AI finances the bridge with perfectly calibrated tolls, and has chosen the contractor and negotiated the price of construction, it will be hard to overcome the desire of humans and their politicians to build the bridge somewhere else and to provide jobs to favored constituents. It is fanciful to imagine that humans can overcome these familiar weaknesses by pre-committing to delegate these decisions to AI, just as (outside the world of science fiction) it is hard to imagine delegation of war-making authority to AI. At best, AI can be a consultant. It is plausible, as we have seen, that even AI would see that combining its skills with a human's skills is superior. Our intuition is that this sort of tilt away from AI is more likely for spending decisions than for regulations, for the reasons already suggested, and also because of the interest of politicians in making irreversible decisions.[12]

C. The Role of Machines

A machine that is engaged in supervised or unsupervised learning[13] adheres to precise instructions for information processing in order to predict an outcome regarding a question that a human has previously defined. It is important to emphasize that AI, and thus machine decision-making, is hard-coded with rules. The very process of assigning a standard or a discretionary decision to a machine, requires detailed rules;[14] true standards

[12] *See* Saul Levmore, *Interest Groups and the Durability of Law in* FRANK FAGAN & SAUL LEVMORE (EDS.), THE TIMING OF LAWMAKING 171 (2017) (suggesting that spending on projects, including bridges and pyramids, reflects a kind of decisionmaking that future lawmakers are unlikely to undo).

[13] *See supra* note 1.

[14] Hard-coded standards, for instance, might consist of a long series of if-then decisions where only a few of those decisions are required to be explicitly "well-reasoned and considered" within the final machine-output decision.

are for human decisionmakers. The same might be said about the ways in which humans make decisions when instructed with standards, but the critical difference is that human judges can rely on intuition and, in any event, (like machines) may be unable to identify their own basis for decision-making.[15]

1. Overstated Concerns with "Unsupervised" Machines and "Opaque" Algorithmic Decisionmaking

Note that unsupervised learning does nothing more than permit the creation or refinement of a rule without further consideration of the machine's accuracy, and perhaps without later supervision or reinforcement of the machine's success or failure.[16] Unsupervised selection of a permanent associate from a pool of two summer associates at a law firm, for example, might be based exclusively upon the patterns in performance over the course of the summer, the patterned performance of past summer associates, and perhaps the harshness of those who evaluate them. If the hiring decision is handed over to AI, it might do more; whether supervised or unsupervised, it might for example include information about the subsequent performance of past summer associates. Law firm partners are unlikely to evaluate their own track records of identifying promising associates, but that is primarily what AI can add with machine learning, so long as it has access to data about future performance and uses it to assess the accuracy of its past decisions.[17]

---

[15] *See* Linda L. Berger, *A Revised View of the Judicial Hunch*, 10 L. COMM. & RHETORIC 1, 1 (2013) (arguing that judicial intuition applied to routine cases often generates bias, but when applied to new problems that require creativity, it provides important solutions).

[16] AI researchers can of course evaluate the performance of unsupervised methods such as clustering, but true unsupervised learning is not evaluated. *See supra* note 1. In some cases, unsupervised learning finds associations or connections which are tagged, and subsequently used as input in supervised learning. The performance of the unsupervised algorithm is then assessed by how much it improves the accuracy of the supervised one. *See* George E. Dahl, Ryan P. Adams & Hugo Larochelle, *Training Restricted Boltzmann Machines on Word Observations*, arXiv: 1202.5695v2 (2012) (applying this approach to analyze word representation).

[17] Consider that COMPAS, a leading AI tool for predicting recidivism, collects self-reported data from defendants. State v. Loomis, 881 N.W.2d 749, 753 (Wis. 2016) (noting that "[t]he COMPAS risk assessment is based upon information gathered from the defendant's criminal file and an interview with the defendant"). If defendant self-reporting is systematically biased, then even though the data is labeled by defendants, predictive accuracy may be *increased* by using some form of unsupervised learning, such as clustering, to find incorrectly labeled patterns that predict recidivism.

This might be an example of isolated work completed outside of the combination of Human and AI, or it might have taken place because of good joint work; perhaps a human thought of the value which would be added by AI and therefore provided the AI with information about future job performance in the years after associates summer at the firm. Machine learning can in this way be more rule-based than is generally recognized,[18] and even unsupervised learning relies on identifying data for collection. If truly unsupervised learning is problematic, because of its opacity, it is because it ignores the role that humans play in goal-setting and identifying appropriate sources of data for collection. Algorithms that allocate the seats of an incoming freshman class on the basis of zip codes, for example, may indeed generate biased outcomes, but it is the human who permitted the data to be collected in the first place.

2. Machine vs. Judicial Decisionmaking

Consider an algorithm that predicts whether a Delaware judge will pierce the corporate veil. This algorithm might collect variables such as the presence of fraud, a bankrupt subsidiary, and the level of capitalization relative to debt. Using those variables as inputs, the algorithm provides detailed information on the expected outcome of a piercing decision and then makes a suggestion.[19] When judges apply rules, they engage in a similar practice. Rules circumscribe their consideration of particular facts and circumstances.[20] While judicial interpretation of both facts and rules can loosen that restriction, standards permit greater freedom in decision-making.[21] In this sense, the application of a standard has more in common with decision-making by the combined effot of humans and a machine—as judges are paired with AI. Of course, in most cases this intelligence is not what we commonly call AI, but is rather the intuitions and conclusions reached by earlier lawmakers who observed some data and offered some imprecise guidance, outsourcing decisionmaking to future judges. This form

[18] CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION (2016); Brent Mittlestadt et al., *The Ethics of Algorithms: Mapping the Debate*, 1 BIG DATA AND SOCIETY 5 (2016).

[19] In this case, the machine would be engaging in reinforcement learning if it received output data, such as aggregate corporate debt for one year following the piercing decision, and used that data when making future decisions.

[20] Isaac Ehrlich & Richard A. Posner, *An Economic Analysis of Legal Rulemaking*, 3 J. L. STUD. 257, 258 (1974) (noting that the difference between a rule and its meta standard is usually one of degree, just as most standards lie below yet more general aims, or standards).

[21] *See* Cass R. Sunstein, *Problems with Rules*, 83 CAL. L. REV. 953, 1021-22 (1995) (noting that rules can never eliminate interpretative uncertainty, but they nonetheless reduce discretion).

of combined power is quite similar to that we have associated with Human-AI combinations. It makes little difference whether a modern judge uses insights offered by a modern machine, a standard constructed by an earlier lawmaker, or indeed an earlier precedent that can now be overruled. In all these cases, the combination might or might not reveal the "reasons" for its suggestion, and the current judge is empowered to use his or her thinking to decide how much to rely on the available machine.

3. Reinforcement Learning and the Importance of Measuring the Performance of Human-AI Combination

It is apparent that for AI to know the right balance between machine and human decisionmaking, it is necessary to have good data about the performance of both humans and machines. The same can of course be said about precedent constructed by earlier humans. Thus, it is important to track the decisionmaking of judges and regulators, in order to compare their judgments with eventual results. This requires far more data collection than is currently practiced, with an eye on both structured and unstructured analysis. The potential advantages of AI require good big data. Just as we ought to collect more information about the classroom, to see what works in the world of education, so too we need to gather information about judges and legal rules, and then the outcomes generated after the application of standards and rules used by law.

Detailed measurement of judicial performance can identify talented judges who have good track records of successfully overruling machines (a word we might begin to use to include earlier judges). Ideally, the judge's history of overruling will itself be examined by AI and will inform this and other judges whether overruling has been done too frequently. The process is improved if the judge is able to explain, at the time of the decision, why overruling seemed right. It is also improved if data about performance can be set aside, and then examined; after all, *some* judge will look good merely by chance. By analyzing their decisions in detail, law may make progress in identifying the true sources of good judicial decision-making, and this can be useful in instructing other judges. Data collection can in this way lead to the refinement of a standard or the creation of a new rule.[22]

III. OVERRULING AI

[22] *See infra* Section IV.B.

It should be noted at the outset that the attention we now draw to the problems of overfit, omitted variables, and reversals, do not offer a comprehensive means of evaluating the optimal balance of power in the combination of a human with AI. Humans, and especially judges, are known to suffer from a variety of decisionmaking flaws that are difficult to overcome even when brought to the attention of decisionmakers. Judges and other lawmakers (and we) suffer from hindsight bias, unconscious biases, and many more defects that AI might avoid—though we should not ignore the impact of biases on the information (the goals as well as classes of data) given to the AI by humans.[23] The issues we raise here are meant to draw attention to the limits of AI, but we are not suggesting that humans are perfect decisionmakers.

A good question to ask at this point is what level of confidence should be required before decision-making is entirely outsourced to the machine or before a human overrules the AI. In some situations, the second version of the question is easy. Imagine that an AI makes a startling recommendation and a human thinks it can identify the source of the AI's surprising decision. Perhaps the AI says to drive at 100 miles per hour because it is clever enough to see not only a safe stretch of road but also that it will make better decisions in the future if it builds up data about safe driving speeds.[24] Whether or not the human understands the source of the AI's astonishing instruction, the human might appreciate something not previously included in the set of goals given to the AI; the human will suddenly see the likelihood that an accident at such high speed will cause voters to distrust or de-fund AI. The human knows that an accident at a speed no human recommends would bring out the worst in voters and consumers, who would not be satisfied with a developer's promise to improve the AI's instructions. Humans are unlikely to remember all the times that AI lowered the speed and saved property and lives, just as they presently respond to a

[23] *See, e.g.*, Jeffrey J. Rachlinski et al., *Does Unconscious Racial Bias Affect Trial Judges?* 84 NOTRE DAME L. REV. 1195, 1195 (2009) (documenting implicit racial bias in the decisions of trial judges); Jeffrey J. Rachlinski, *A Positive Psychological Theory of Judging in Hindsight*, 65 U. CHI. L. REV. 570, 571 (1998) (asserting that judges and juries overstate the predictability of past events); *see also* Daniel L. Chen & Jess Eagel, *Can Machine Learning Help Predict the Outcome of Asylum Adjudications?*, Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law 237, 237 (2017) (documenting judges' past decisions to grant asylum as predictors of future decisions).

[24] Humans expect an increasing function, while the AI might see that the truth may be otherwise. The incongruity would be avoided if a human knew to set a goal for the AI that included the risk that error would lead to pushback. Thus, the AI might be told to see what minimizes serious accidents over a weighted ten year period or to weight endangered parties' lives in a way that accounts for their apparent ages.

reported accident involving an autonomous vehicle without properly weighing all the accidents that humans cause and AI avoids.

But now consider overrides in the other direction. Imagine a case where the human lawmaker insists that a factory should be required to mitigate pollution in some manner, but the AI partner surprises by suggesting that the polluting activity go unmitigated and that, instead of adhering to previously established environmental law rules, the factory owner (or government) should pay two other factories on the same river to operate fewer hours a day so as to improve the ecosystem at lower cost. Put differently, the AI sees a Coasian bargain where transaction costs or a failure of imagination inhibited human regulators and social progress. How certain should a human lawmaker be that the AI is wrong before intervening? The human may calculate that the AI is more likely than not to be wrong; perhaps it seems incorrectly instructed about the costs and benefits of displacing employees in closed factories. An important literature suggests that a more-likely-than-not, or 51% certainty, is error minimizing and efficient; the negligent or strictly liable defendant should pay full rather than probabilistic damages.[25] But now the AI has presumably used big data to calculate costs and benefits, and it has determined that there is a 52% chance, let us say, that an exception to existing environmental law is appropriate. In turn, the human regulator has calculated that with similar probability the AI has incorrectly assessed the true costs of change. Our claim is that the AI should be trusted (and not overruled) more readily than the human when humans are confident in the precision of their specification of AI's goal (as discussed in Section II.C.1 above); when the legal domain is sufficiently stable over time and new variables are unexpected to appear; and, where causal models draw on AI to identify important variables, that these models are robust to data division, especially when omitted variables may be a problem. Our concern with stability is reflected in errors generated by overfitting a predictive model. Concerns with robustness are reflected in errors generated by omitting important variables (or by misspecification generally) in a causal model. We take up both of these in turn.

A. Overruling On the Basis of Overfit

AI's accuracy depends upon whether the relationship or pattern that it has identified exists in the real world. From the outset, it is important to

[25] David Kaye, *The Limits of the Preponderance of the Evidence Standard: Justifiable Naked Statistical Evidence and Multiple Causation*, 7 AM. BAR FOUND. RES. J. 487, 487 (1982); Saul Levmore, *Probabilistic Recoveries, Restitution, and Recurring Wrongs*, 19 J. LEGAL STUD. 691, 691 (1990).

note that contemporary machine learning (ML) is generally used to make predictive inferences; it might for example indicate that the presence of characteristics X, Y, and Z predicts that the bailee is a flight risk. A policy which eliminates characteristic X from the population, say, by increasing everyone's income by $1,000 per month, may or may not cause a given defendant to be marked as someone who presents a low risk of flight from the jurisdiction. Researchers are actively working toward supplying AI with the tools of causal reasoning to reach stronger conclusions.[26]

In statistics, randomized experiments are the gold standard for identifying a causal relationship.[27] Random sampling from a population, and random treatment across a subset of that sample with a drug or policy, for example, sharply reduces omitted-variable bias.[28] The challenge for law is that random experiments are rare; jurisdictions create rules for reasons, and their random application, even within jurisdictions, is scarce if not legally risky.[29] The implication is that the traditional approach of increasing the number of independent variables to combat omitted-variable bias does not translate well to law. As we will see, there remains a worst-case scenario where a causal inference is not only wrong, but the *reverse* of what is empirically apparent.

Predictive models promise to alleviate the drawbacks of causal inference and omitted-variable bias, but they face their own challenge of "overfit." Indeed, we might say that all searches of multiple variables for a

[26] *See* Judea Pearl, *Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution*, arXiv: 1801.04016v1 (2018) (noting that advances in graphical and structural models have made counterfactuals computationally tractable and that machines can be equipped with the tools of causal reasoning).

[27] *See generally*, RONALD A. FISHER, THE DESIGN OF EXPERIMENTS (Hafner Publ'g Co., 6th ed. 1951) (1935).

[28] Omitted-variable bias occurs when a relevant variable that actually belongs in the true population model, and is correlated with an identified regressor, is excluded from the statistical model hypothesized by the researcher. Because the variable is excluded, the model is said to be "underspecified." Any estimates of the magnitudes or signs of the included variables may be incorrect, or "biased" as a result. Properly randomizing treatment application across a sample does not reveal omitted variables, but it ensures that the treatment and omitted variables are uncorrelated, which ensures that omitted-variable *bias* does not arise in turn. *See* JEFFREY M. WOOLRIDGE, INTRODUCTORY ECONOMETRICS: A MODERN APPROACH 88-92 (5th ed. 2013).

[29] *See* Michael Abramowicz, Ian Ayres & Yair Listokin, *Randomizing Law*, 159 U. PENN. L. REV. 929, 964-74 (2011) (discussing the potential need for informed consent for inclusion in a legal experiment and possible equal protection violations for those who are excluded from the random application of a rule). On randomized law's uneasy relationship to equal protection, see also RONALD DWORKIN, LAW'S EMPIRE 178-84 (1986) (discussing problems with arbitrariness).

best fit suffer from the danger of overfit, but that the risk of AI in this regard is likely greater than that faced by human data analysts because the latter begin with theory and look at few variables.[30] Machine learning primarily relies on raw data, as opposed to theory, to construct models and generate testable hypotheses. Because its models are presently data-driven, it faces a tradeoff between demanding that new observations for classification or prediction perfectly match all of the features of the existing data set from which a model is constructed.[31] For instance, a model may only identify a Coasian bargain for mitigating pollution if a factory shares identical features of all previous factories where Coasian bargains made sense. This form of algorithmic rigidity, or overfitting of the data to the model, may hide other worthwhile bargains from the policy-maker's reach. On the other hand, if an algorithm permits too much dissimilarity, it may erroneously classify a factory as ripe for a bargain, when a bargain, for a factory with these additional unconsidered and unmodeled features, actually leads to social loss.

1. An Example of Overfit with Respect to Judicial Precedent

Most lawyers are familiar with overfit, but perhaps not by that name. The use of precedent can be understood as the judicial strategy, or restrictive principle, of fitting a classification rule to the evidence. Common law rules are developed and refined over time by the presentation of various configurations of facts and circumstances to courts. The facts and circumstances are analogous to training data that generate a model rule, which a court applies to a future set of facts and circumstances in order to rule on and "classify" a case. A correctly decided case strengthens the classifier and increases the likelihood that a new case will be properly classified. If a lower court interprets precedent too loosely, and attempts to shoehorn the facts to the rule, then it has overfit the case; in principle, the error is corrected on appeal.

2. Overfit in the Process of Releasing Arrested Felons on Bail

[30] Large data sets enabled by big data collection often contain an extremely large number of variables, and a choice must be made in order to reduce their number to a level manageable for causal inference. A good match can be superior to a perfect one, if the former is located with a theory that proves useful in predicting the future as it identifies causal factors; theory might be robust in the face of change, while AI knows only the past. *See* Pearl, *supra* note 26 at 6 (noting that the problem of robustness in the face of environmental change requires a causal model, which "cannot be handled at the level of Association").

[31] These new observations are held out while constructing the initial model.

In terms of AI, and its ability to make predictive inferences, the problem of overfit is always present so long as an additional observation adds at least one new variable to consider. This can be difficult to understand because new data that confirms a hypothesis normally increases the posterior probability that the hypothesis is correct. If Judge A thinks that every person arrested for a felony should be released on bail only if the bail is greater than $1 million, and someone released by Judge B on $500,000 bail commits a serious crime, A's thinking about the million dollar rule is strengthened; B released the person on bail set below $1 million, and sure enough a serious crime followed. But this new piece of data also, and almost necessarily, introduces a new variable. The most recently observed and released person might have enough money saved to flee the jurisdiction and have his face altered, or he might not have a job waiting for him, or he might have been released on a Wednesday. In principle, any of these factors, or a combination of them, might increase the chance that another hypothesis best fits the available (past and newly arrived) data. It may be that Judge B's judgment is terrible and now it will become evident that anyone whom B wants to release requires bail of $2 million or an ankle bracelet. It may be that a secure job is a better predictor than is the amount of bail because the case sent to B involved an arrested person without a job, and a new study of the data, old and new, now shows that unemployment is a better predictor than the amount of bail, or that job insecurity combined with a $600,000 bail cutoff is the best predictor, and so forth. In short, the new piece of data might raise the posterior probability attached to the one million dollar rule, but it might also require further study because it introduces the possibility that another variable is more important than the amount of bail. If the goal is precision about preventing crimes and flight, then even as the algorithm benefits from new data, it loses in the sense of needing to examine or re-examine other variables. A new observation can make the old algorithm look worse for obvious reasons, as it would if Judge B set bail at $1.2 million but a crime was experienced, or if B set a low bail and no crime followed. But the more difficult point is that the new observation may appear to confirm the value of the previous rule, but it may also prove to make another rule superior.

This problem, if that is the right word, is not easy to solve, as the machine investigates a larger number of hypotheses (and note that the characteristics of the new piece of data need to be combined with all other characteristics because the winning formula might be something like "bail amount and waiting job and an ankle bracelet and wearing a tie in the courtroom") it locates omitted variables but has more trouble discerning true connections. The problem of overfit is now apparent: with enough

variables and attempts to find a theory, there is a good chance that the best fit was nothing more than a random result.[32] Some combination of features is bound to look like the best predictor of the desired result. Ideally, we would divide the data and see how the best predictor fairs with the set-aside data. But this approach is not ideal for lawmaking because circumstances change. The best predictor of successful release on bail in 1995 might be expected to be very different from the most important variables in 2019. Ideally, we would wait for new data; this comes quickly in some settings, like testing remedies for the flu, but it is impossibly slow for many legal applications. If greater value is attached to more recent observations, then the number of useful observations is reduced. While big data may virtually eliminate the problem of overfit in some legal domains where law is relatively settled, new forms of human activity and its regulation will suffer from limited data, weak models, and overfit. Learning cannot occur, and is certainly less reliable, when the context of a generalization is changing.[33]

B. Overruling AI on the Basis of Omitted Variables

AI and machine-learning applications are generally used for predictive inference, but in some cases lawmakers wish to make causal inferences. If policy X is adopted, will adoption cause outcome Y to occur? As noted above, the challenge of developing accurate causal models involves identification of a model for testing. Presently, researchers are beginning to use machine learning to help predict (and ultimately identify) important variables for causal models and to select correct causal model structures.[34] In this way, the future of ML presents a mix of predictive and causal reasoning that is susceptible to both overfit and omissions.[35] If lawmakers

[32] *See* RUSSELL & NORVIG, *supra* note 1 at 705, which provides the example of attempting to predict that the roll of the die will turn up as 6 by collecting data on the color of the die, the time of day the die is thrown, its weight, and whether the experimenters crossed their fingers when casting. They note "[i]f it turns out that there are 2 rolls of a 7-gram blue die with fingers crossed and they both come out 6, then the algorithm may [erroneously] construct a path that predicts 6 in that case." *Id.* The solution to this danger is, presumably, to try the algorithm on future or set-aside tosses of dice in order to expose, and then correct for, the overfit. Another solution is to trust the initial theory, usually devised by humans, that tosses produce random outcomes, and that each toss of a die offers a 1 in 6 chance of turning up a 6.

[33] *See* VALIANT, *supra* note 9 (discussing the Learnable Regularity Assumption and the Invariance Assumption).

[34] *See, e.g.*, Frank Fagan, *Big Data Legal Scholarship: Toward a Research Program and Practitioner's Guide*, 20 VA. J. L. & TECH. 1, 25 (2016) (noting that machine learning can be used to identify correlations between judicially determined facts and decision-making, which can later be used for causal inference).

[35] *Cf.* Pearl, *supra* note 26 at 6 (noting the necessity of developing causal models when

rely on AI to make causal inferences, even indirectly through AI's assistance in specifying a causal model, then lawmakers should be skeptical when these models fail basic robustness checks, such as split-sample analysis, since failures suggest that omitted-variable bias may be present.

1. The Omitted Variable Problem in Bankruptcy

Bankruptcy law offers one area in which the omitted variable problem is easy to see. Suppose that a causal model shows that an absolute priority rule, which can be overridden with a majority vote, preserves the largest percentage of wealth of a debtor's estate when contracting after bankruptcy is barred.[36] Imagine further that a second study examines two groups of debtors: group A consists of debtors with more than six tranches of credit, and group B with fewer than six. When examining groups A and B together, a majoritarian absolute priority rule, which bars ex post contracting, maximizes the wealth of debtors, and has no unattractive third-party effects. However, if the groups are evaluated separately, then the relationship reverses for both A and B; a majoritarian absolute priority rule which permits ex post contracting maximizes wealth.

What might cause such a reversal in this center-piece of bankruptcy law? The most obvious reason is that the first study failed to include a categorical variable which would have captured the difference above and below a certain number of creditors. Earlier work has shown that information asymmetry between corporate managers and creditors leads to deviations from the absolute priority rule.[37] Perhaps the presence of fewer tranches is correlated with greater asymmetry and deviations, inasmuch as dishonest managers can only fool a few creditors. In turn, greater asymmetry and deviations are negatively related to wealth irrespective of permissible ex post contracting. When ignoring the tranche threshold, information asymmetry between corporate managers and creditors is effectively ignored, and causes the reversal.

---

machine learning is not robust to changing conditions).

[36] The Bankruptcy Act of 1898 mandated an absolute priority rule and provided for no majority override. The Act was amended in 1978 to permit majority override, but post-bankruptcy contracting still remains controversial. On the history of the Act, see DAVID A. SKEEL, JR., DEBT'S DOMINION: A HISTORY OF BANKRUPTCY LAW IN AMERICA (2003). For a discussion of the merits of ex post contracting, see David A. Skeel, Jr. & George Triantis, *Bankruptcy's Uneasy Shift to a Contract Paradigm*, U. PENN. L. REV. (forthcoming 2018).

[37] Maria Carapeto, *Explaining Deviations from Absolute Priority Rules in Bankruptcy*, 3 J. EMPIRICAL LEG. STUD. 543, 555 (2006) (providing evidence that when payments in equity (a proxy for management's overstatement of firm value to creditors) constitute a larger proportion of total consideration, deviations from absolute priority tend to occur).

A less obvious reason is that the first study may have identified a statistically significant relationship, but only for a fraction of the debtors' estates observed. We can think of the majoritarian absolute priority rule (barring contracts after bankruptcy) as a "treatment." Perhaps out of a total of 5,000 observations, there are 1,000 observations of debtors' estates where this treatment is applied. Of those 1,000 observations, 550 experience an unambiguous increase in wealth relative to the 4,000 observations which are given no treatment. In other words, the treatment increases wealth 55% of the time within the observable data. Even if the relationship between the treatment and this relatively frequent increase in wealth is statistically significant at conventional levels, the empirical result suggests fragility. The greater the number of observations and the higher the percentage of observations showing an increase in wealth, the more confident we can be that the connection is noteworthy and not fooled by an omitted variable.[38] If 45% of the estates experience no increase in wealth following treatment, then evaluation of an additional variable, such as the number of creditors above or below six, might lead to empirical indeterminacy, or suggest an alternative rule. Consideration of this additional variable effectively unmasks what dividing the data into subgroups for validation reveals, namely the importance of creditor structure and information asymmetry for determining the rule's effectiveness.

A reversal is dramatic. It may be that dividing observations into subgroups, or adding variables, only leads to indeterminacy.[39] When the statistical significance of correlations and causal relationships merely collapses as the result of additional empirical study, no new rule is suggested and law might simply go unchanged. But suppose further study considers the number of creditors as a continuous variable (as opposed to a discrete variable, indicating whether the number of creditors is greater or less than six), and it also fails to reveal a statistically significant relationship between ex post contracting and the wealth of a debtor's estate. This result would suggest that the second empirical study is fragile, and that the relationship between creditor structure, information asymmetry, and the timing of contracting is not sufficiently understood. Plainly, some important variable remains undetected or incompletely measured. Law may do well to respond here with a standard; the standard would allow the bankruptcy

[38] On the relationship between the strength of the effect under consideration and the confidence one can place in the results of significance testing, *see infra* note 47, particularly the discussion of "statistical power."

[39] *See infra* Section IV.C.

judge to decide whether ex post contracting should be permitted.[40]

2. The Limitations of Using Machine Learning to Detect Omitted Variables

A machine learning algorithm may fail to detect an important variable if there is too little data to distinguish among candidate hypotheses. A conjunctive machine-learning algorithm, for instance, formulates a new hypothesis each time the data accounts for an additional variable; an absolute priority rule's relationship to creditor joint-wealth may depend on the number of tranches, the presence or absence of an equity trance, the proportion of payments in equity, the relative magnitude of senior to subordinate tranches, the jurisdiction of the bankruptcy, and the tenure of the current board. The number of possible hypotheses grows exponentially for each additional variable observed.[41] Suppose the AI draws each hypothesis at random and resolves to keep the one that best matches the data. While more data would reduce the number of hypotheses that survive, some will still remain if an analysis begins with bad hypotheses. In principle, the AI can compute the amount of data needed to approximate a correct hypothesis, but in practice the amount of data will be limited.[42]

For this reason, AI tests the accuracy of its chosen hypothesis on new data that is set aside during its initial development. Researchers often subject identified patterns and superior hypotheses to significance testing in order to combat overfit.[43] A hypothesis that identifies 800 increases in wealth versus 200 non-increases is probably accurate, but not if it is the best one among several thousand hypotheses. Achieving 80% accuracy once in several thousand attempts can easily occur by chance. Some false hypotheses will inevitably be accepted, but by rejecting statistically insignificant ones and testing those that remain with new data, overfit can be controlled. (Another method is to prefer simpler hypotheses.) If the model continues to fit the data too tightly and the same mistakes continue to surface (with bias), the combination of Human-AI can introduce more flexibility through overruling or reducing the required threshold for significance.

---

[40] *See infra* Section IV.C.

[41] *See* PEDRO DOMINGOS, THE MASTER ALGORITHM 73 (2015).

[42] *See* VALIANT, *supra* note 9 at 74 (explaining that the number of hypotheses usually exceeds the number of observations since new observations introduce new variables that expand the hypothesis set).

[43] *See* DOMINGOS, *supra* note 42 at 73-74.

C. Overruling AI on the Basis of Type II Errors

1. An Example of Type II Errors in Veil-Piercing Corporate Law

Consider an empirical study which shows that a judicial finding of undercapitalization has no statistically significant relationship with a judicial decision to pierce the corporate veil. Suppose that other variables—such as categorical variables reflecting the presence of fraud, bankrupt subsidiaries, or subsidiaries that can avoid statutory obligations such as CERCLA liability if the assets of the parent remain inaccessible—are statistically significant at the conventional 5% level.[44] This means that there is a 5% chance of observing those relationships in the data even if those relationships do not exist; we may be fooled by randomness. In principle, the relationships may be real and could be expected to show themselves again in an overwhelming number of repeated inquiries.

There are at least two reasons why these results may, nonetheless, fail to replicate. Consider first the lack of evidence of undercapitalization's significant effect. A study that follows convention, and sets the threshold for statistical significance at 5%, will fail to identify a relationship between undercapitalization and veil-piercing if the data reflects just a 6% chance of observing that relationship at random. This means that there is a 94% chance that the relationship could actually exist.[45] For this reason, as the threshold for statistical significance increases, the possibility of failing to reject the hypothesis that undercapitalization has no effect when it actually could have an effect (a Type II error) increases. If the threshold for failing to reject the relationship is set to 1%, but the data reflects just a 2% chance of a random relationship, then there is a 98% chance that a similar study will find that the relationship between undercapitalization and veil-piercing could actually exist. Even if predictive models are unconcerned with significance testing on their own, researchers use testing and other rules of exclusion to select among candidate models and compare performance.[46] The danger of failing to exclude an incorrect, or less accurate, model thus

[44] *See* Jonathan Massey & Joshua Mitts, *Finding Order in the Morass: The Three Real Justifications for Piercing the Corporate Veil*, 100 CORNELL L. REV. 99 (2014) (providing empirical evidence that fraud, bankruptcy, and the advancement of federal statutory purpose predicts a judicial decision to veil-pierce, but that a factual finding of undercapitalization does not).

[45] More precisely, if the researcher's initial belief before the study (the alternative hypothesis) is that undercapitalization predicts veil-piercing, failing to reject its converse (the null hypothesis) does not imply that undercapitalization does not predict veil-piercing. The study simply fails to reject the null.

[46] *See* DOMINGOS, *supra* note 41, at 87.

remains.

2. The Severity of Type II Errors in Law

The problem just discussed can be particularly severe in law, and it suggests a significant role for lawyers and lawmakers. Apart from its importance in coordinating society, enabling transactions, and creating Pareto-improvements, much of law draws a sharp line between winners and losers. A rule which permits contracting after bankruptcy favors some class or classes of creditors at the expense of others. A plaintiff may have no claim against an undercapitalized defendant if undercapitalization were statutorily abolished as a veil-piercing rationale. In contrast, a drug found to be ineffective during trials does not suggest that patients should be treated with another drug, readily at hand, or left to suffer with no drug at all. Moreover, medical treatments are not mutually exclusive; in law, on the other hand, mutual exclusivity is the norm as in the case of speed limits or a rule permitting or forbidding ex post bankruptcy contracting. If a medical treatment is effective for only 55% of the population, the other 45% does not suffer from submitting to that treatment, absent side effects. In any case, the patients still undergo different and additional treatment if the initial one fails.

This difference between law and medicine is systematic, and the argument in this Article is not simply a matter of picking favorable examples to justify its claims. In law, rules that simultaneously assign mutually exclusive rights, normally create winners and losers. While some medicinal treatments may be mutually exclusive because of finality (amputation) or time (daily travel to a treatment center, or taking drugs with short-term effects), these are exceptions. Most treatments are additive. Legal rules that favor one group at the expense of others, or that simultaneously permit and forbid behaviors, are particularly susceptible to reliance on flawed empirical research – because of predictive inaccuracies or causal errors. There is a cost to permitting person A to do something that is socially costly, but there is an additional cost if the same rule forbids person B from doing something that is socially valuable. When one group receives an unwarranted benefit at the expense of another, the rule does not simply reward members of the first group; it makes the second group worse off. Empirical missteps in law are in this way often amplified.[47] The same is

[47] Fashioning rules on the basis of Type II errors, or failures to reject false null hypotheses (such as undercapitalization does not predict veil-piercing) when those hypotheses should be rejected, is problematic. It is often thought that Type I errors can be limited and controlled by strengthening the significance threshold. If, for instance, the

true for expenditures, though the negative effect is often dispersed and unnoticed; building a bridge in the wrong location but in favor of one interest group, usually affects another, either because its members must now travel far to the bridge or simply be taxed for construction that benefits them not at all.

D. Overruling or Preferring AI because of Simpson's Paradox

The problems posed by omitted variables is even worse when examination of the data is susceptible to "reversal paradoxes," for then an inference is not just inferior to other techniques, but it is *very* wrong. In these instances, law will do well to abandon its reliance on rules, and instead prefer standards, or other forms of flexible decisionmaking.[48]

A reversal, or Simpson's Paradox, generally requires a different distribution of observations in the sets of data that are compared; it is more likely to surface when there is a small number of observations and when the degree of certainty about a conclusion is low. For these reasons, it is more likely to arise in causal than in predictive inference, though predictions of outcome A may reverse to B (or not-A) if a predictive model is susceptible to overfit or omissions.

1. An Example of Reversals in the Context of Evaluating a Law Firm's Summer Associates

---

relationship between an event, such as the announcement of a merger, and increased stock price is significant at the 5% level, we can be even more certain that the observed relationship is not a product of random chance, by tightening the significance threshold to 1%. As we have seen, alteration in the threshold does not come without its cost, as it increases the study's susceptibility to Type II errors. Suggesting the *absence* of a causal relationship between the merger announcement and increased stock price, even at the 1% level, could be erroneous on its own if the study lacks sufficient statistical power. "Power" is the probability of correctly identifying an effect within a population. Statistical power decreases with smaller significance thresholds, sample sizes, and effects under investigation – that is, the strength of a relationship between two variables. Of these three, effect size introduces the greatest uncertainty in a study. If the price impact is small, an inference that the announcement caused such impact, even if statistically significant, could be erroneous. *See* Jill E. Fisch, Jonah B. Gelbach & Jonathan Klick, *The Logic and Limits of Event Studies in Securities Fraud Litigation*, 96 TEXAS L. REV. 552, 618 (2018) (noting that a significant effect between events and changes in stock prices may go undetected if that effect is small); Alon Brav & J.B. Heaton, *Event Studies in Securities Litigation: Low Power, Confounding Effects, and Bias*, 93 WASH. U. L. REV. 583, 600-01 (2015) (same).

[48] On standards, *see infra* Section IV.C. On other forms of flexible law, see, e.g., Frank Fagan, *Legal Cycles and Stabilization Rules*, in FRANK FAGAN & SAUL LEVMORE (EDS.), THE TIMING OF LAWMAKING 11 (2017) (discussing contingent rules).

Imagine a law firm looking to make an offer to one of two summer associates, Kim and Kit. The firm decides to score the associates on assignments given to them while they summer at the firm. During the first month of the summer, Kit is given one corporate assignment and deemed to have done a poor job on it. Meanwhile, Kim is given four assignments in that department and is graded as a success on one of them. During the second month, the two are assigned to the environmental group. Kit receives five assignments and succeeds on four. Kim is deemed to have successfully completed the two assignments given in the same department. The firm tabulates the reviews and decides to hire Kit because Kit impresses on 4 of the 6 assignments, while Kim impressed partners on just 3 of the 6. But then a partner points out that perhaps the corporate assignments were simply more difficult, or the corporate partners graded more ferociously than those in the environmental group. Indeed, Kim performed better than Kit on the corporate assignments, and *also* better than Kit on the environmental projects. Each department would prefer Kim over Kit, even though Kit's overall score was superior. This is a classic reversal paradox; in this example the reversal easily came about because the summer associates were not (and probably could not be) given the same assignments, and not even the same number of assignments in each department. Note that a further, or double, reversal paradox is possible. Perhaps partners tend to give high scores in the morning, and Kit was always evaluated in the afternoon. Had they both been evaluated in the morning, or afternoon, Kit would have been thought superior in both departments, and on *both* morning and afternoon assignments.

Even if Kim and Kit had been given the same assignments, from the same partner and at the same time of day, reversal would continue to be a danger because of their dissimilar innate abilities and characteristics. Kim may be bilingual and able to complete immigration cases more quickly than Kit. Kit may be better at managing work-life balance and coping with stress than is Kim. Any model that omits variables that account for those dissimilarities is systematically biased and consequently fragile. No single variable describes successful completion of a law firm assignment. Thus, one evaluator may consciously or otherwise reward good writing, while another is drawn to the summer associate's ability to produce a memo quickly. Such differences can bring about a reversal paradox when partners simply use a generic variable like "ability," and make future success less predictable, until specific variables are identified.[49] The disparate outcomes,

[49] Even with a specific variable, reversals are possible. For example, a coach may like speed in a race, but one runner may be faster when the running path includes hills, while another wins when the weather is between 60 and 65 degrees. As described presently in the

even when Kim and Kit are given an identical task in an identical manner, reveal "heterogeneous treatment effects," and these effects across people, business entities, or other objects of study, can generate reversals, as they are a manifestation of omitted variables. It is simply the case that hidden variables can bring about reversals.

It is common to rely on empirical evidence but to question researchers about omitted variables. When Kim is preferred after the initial analysis of performance, someone whose intuition was to favor Kit might have pointed out that Kit's assignments were more difficult or that Kit was evaluated by partners who tended to be tough graders. If the omitted variable—difficulty—were properly included, the result would have been different. But in most cases this would mean that one department would favor Kit over Kim or that the overall scores would be different. The remarkable thing about the special case of a Simpson's Paradox is that an omitted variable causes *both* departments to favor one result, while the overall, combined score still favors the opposite result, even when all known variables are included. The practical and often startling lesson is that even when an empirical study is questioned because of some omitted variable, and that variable is included in further study, the result may still favor X over Y, even though Y is superior to X in every setting. This will not occur if one knows exactly how to weight the omitted variable, but the weight itself is often unknown or unmeasurable, and can, in any case, be thought of as an additional omitted variable. Assigning a weight or properly modeling a relationship is often difficult, and it is hard or even impossible to know when it has been done correctly.[50] Thus, in the law firm case, if one partner suggests that Kit may have been given more difficult assignments, and Kit's scores are therefore increased by some percentage, a reversal paradox may yet occur because the increase was imperfectly sized, and lower than required to escape the paradox.

Reversal paradoxes can be understood as a subset of the problem of omitted variables, but it is a particularly interesting subset both because of the startling reversals and because these problems are more difficult to solve than the mere additional testing that can respond to most skepticism based on omitted variables. Dividing and validating data is a practice that avoids many problems in empirical work, and it can reduce the risk of reversals. If

---

text, hidden variables are the stuff of reversals.

[50] While large data sets and reliance on prediction can lessen the need for causal inference, empirical models which identify the structure of a predictive relationship with big data can suffer from overfit; the model may correspond too closely to the data and fail to reliably predict future observations as a result. *See supra* Section III.A.1.

there are 20,000 patients with a disease and the scientist wants to test combinations of drugs, it is usually wise to find the winning combination on a group of 10,000 randomly chosen patients, and then test this finding by applying it to the previously untested, or set-aside, 10,000. In our necessarily (and unfortunately) smaller case, if Kim and Kit are evaluated over two summers, rather than one, and Kim earns higher total scores in both summers, it is less likely that Kit was unknowingly given more difficult assignments or graders in each of the two summers. It would probably be a mistake to sum the scores over both summers as one set of data, because a hidden reversal paradox is more likely in one comparison than in two. On the other hand, data division is more attractive when there are many observations, and there are already so few in the law firm example that a division of the two sets of data into tiny sets is likely to make the decision less rather than more reliable. Moreover, a conventional division would hardly be random. Data division, followed by validation, is almost always a good idea when "big data" are available;[51] a random division of the data has become a best practice, although modern techniques often encourage a division that leaves more recent observations in the second data set, where the lessons derived from the first are tested. After all, we are usually eager for results that will work in the future.[52] We can think of this strategy as selectively random. Another solution to the reversal paradox problem might be to increase the number of observations. If the assignments are broken down so that Kim and Kit are scored on 20 assignments rather than 6, it is more likely that the difficulty factor averages out.

2. An Example of Omitted Variables and a Reversal with Respect to Autonomous Vehicles

It is apparent that reversal paradoxes can occur in the context of comparing AI and human performance and thus make it difficult to know when the AI should be overruled. Our discussion therefore turns to a few examples of the paradox in this context. We begin with the now familiar question of Human-AI interaction with respect to autonomous vehicles, and then imagine cases that will arise in making law.

---

[51] Predictive models are generally trained and tested with set-aside data. Current techniques, such as k-fold cross-validation, divide the data into a number of random partitions for estimating model performance; accuracy rates are assessed for each partition, or fold, to determine adjustments to the model. *See* BRETT LANTZ, MACHINE LEARNING WITH R 319 (2013).

[52] Note that this approach essentially omits observation of time-dependent effects, though perhaps the bigger pitfall with data-generated models is overfit, described *supra* in Section III.A.

Consider the case of a human investigator who has reason to think that it is 60% likely that an accident involving an autonomous vehicle, relying on AI, was caused by the AI's error. One possibility is that the misstep might really be blamed on a human's failure to take an outcome, or goal, into account in programming the AI in the first place. Perhaps after the AI and automobile have been in use for some time, the human considers 100 accidents that have occurred and sees that in 60 of these it would have been better for the AI or a human driver to take a step contrary to the one recommended or carried out by the AI. For example, the AI may have brought the vehicle to a halt, when blowing the horn would have saved the day. Imagine further that the AI does not appear to have improved over time. Simpson's Paradox warns us that the better result might be the opposite of what the human's empirical evidence suggests. But is the human or the AI more likely to err? Arguably, the AI is less likely to fall prey to a reversal paradox error because it might have tested right away for the importance of many or all avoidance techniques, whereas the human appreciated the value of horn-blowing only through an after-the-fact consideration. It is not enough to say that we can allow the humans and machines to overrule one another and see which does best in practice over a long period, because they might face different problems and the difficulty factor might, once again, be different on the two sides. Even if we wait for a large sample and divide the data, there could be a reversal within some categories, or subsets. It may turn out that humans should overrule more readily in the morning, or that machines are better when the other vehicle is red. Humans must feed data and there is a limit to what data is offered to the machine, however superior it might be. Just as the AI does not always "tell" us how it makes a decision, so too the human may be unable to articulate the source of its "intuition."

IV. LAW AFTER AI OR HUMANS OVERRULE

A. Bigger Data, Deeper Analysis

There are two reasons to think that AI is less likely than humans to be fooled by reversal paradoxes and statistical errors. The first, as suggested earlier, is that AI can handle big data. Reversals are more likely to occur when there is a high level of variance regarding the key variable under scrutiny in the underlying population. In the presence of substantial variance, a random sample is more likely to miss critical features of the population. The initial inquiry generates a biased result when examining the relationship among variables. With a large data set, the data can be divided

into groups and a result tested on the previously excluded data. Once an omitted variable (such as difficulty or time of day in our previous examples) has been identified as a critical factor through division and re-testing, it can be included in future analysis. The larger data set and the inclusion of the additional variable tighten the variance of the estimated variable of interest, and increase the degree of certainty in the statistical result.[53]

The second advantage of AI in this respect derives from its ability to find connections in data; it looks for things that humans do not know, or have not the energy, to examine. AI with big data is constructed to examine the available data (ideally, well collected and accurately labeled) which will include things like (in the case of autonomous vehicles) the sex or height of drivers, weather conditions, left-handed drivers, and car models. Humans, to be sure, must specify the goals; safety is one, but some value needs to be assigned to time (humans like to get places quickly), manufacturing costs, and so forth.

The use of big data, especially when unstructured and not organized by humans, raises the question of whether reversals and other problems are more serious with AI than with humans. AI is able to identify variables that ought to be included in structural modeling, even in the absence of a theory that pointed humans to the variable. In contrast, humans have the advantage of initial theorizing and hypotheses development and these motivate their organization, labelling, and interpretation of data. They combat reversals not by their superiority with data but with knowing what to do with limited data. At times they can gather and investigate more data to improve their conclusions. AI's strength, is its ability to find connections previously unimagined—and yet this is precisely what invites reversal paradoxes. The more startling is AI's discovery, the less it should be trusted, both because it has not been refined and because it may be explained by yet another previously unseen variable. On the other hand, AI is *less* likely to be misled by reversal paradoxes, because it looks for more connections and thus discovers the source of reversal paradoxes that might befuddle humans. Both AI and humans can minimize the reversal problem with the use of big data. To repeat, AI will be guilty of reversal errors because it sees

[53] This is but one of many good reasons to divide data – a practice not yet appreciated in empirical work in law. Here, data division and testing might reveal the significance of new variables. This practice is also relevant for predictive models. *See* DOMINGOS, *supra* note 41 at 75 (noting that hypothesized patterns should be tested on held-out data to combat overfit and that doing so "is just applying the scientific method … to machine learning: it's not enough for a new theory to explain past evidence because it's easy to concoct a theory that does that; the theory must also make new predictions, and you only accept it after they've been experimentally verified").

connections where there is no theory or reason to expect these connections to matter; as a result, it is in danger of missing other variables that reverse the finding. This might be mitigated by dividing data or using other techniques, but the danger remains. On the other hand, inasmuch as AI can consider many times more variables, or theories, than a human, it avoids reversals to which humans will fall prey; AI is more likely to find the very variables that caused reversals in the hands of humans. Our intuition is that the second effect is greater than the first, but further work on this matter is required.

B. Standards to Rules

It may be apparent that AI can turn standards into rules.[54] When it suggests more refined speed limits or divides pollution control among factories on a river,[55] it is turning a standard into a rule, or turning a rule into more refined rules. It promises a more rule-oriented legal system. We have seen that some of these rules might be misguided by reversal paradoxes and other problems. In some cases it might not turn standards into rules; it could, for example, reveal that some or all humans do a better job of predicting repeat offenders or safe drivers than any AI, or an AI developed formula that has been turned over to humans for execution. The AI might show that, for reasons unknown, Judge A is a fantastic decisionmaker when it comes to bail. Both humans and AI might succeed without a capacity to reveal how conclusions are reached. AI might even show that in some areas humans quite generally, rather than AI, are superior predictors and lawmakers. But it is likely that these findings in the direction of standards will be rare compared to the development of new and refined rules.

As we have seen, more data and catalogued experience promise greater

[54] *Cf.* Anthony J. Casey & Anthony Niblett, *The Death of Rules and Standards*, 92 INDIANA L. J. 1401, 1433 (suggesting that extremely detailed rules, or micro-directives, announced at the time that they are needed will emerge as the norm and eliminate both standards and conventional rules, as well as the work of many judges). Note, however, that while revelations of rules improve oversight they also empower those who would evade law. *See* Saul Levmore, *Double Blind Lawmaking and Other Comments on the Formalism of Tax Law*, 66 U. CHI. L. REV. 915, 919 (1999). Simultaneous revelation at the moment of action lowers evasion for a single actor, as a person cannot easily evade an unknown rule, but there remains a social cost to revelation inasmuch as similarly situated persons learn something about the rule's contents by observing its earlier enforcement—even if uniquely tailored by a machine. If actions are heterogeneous, then learning is less costly in terms of evasion; on the other hand, heterogeneity may reduce the social value of a rules-based architecture.

[55] As imagined earlier inSection III.A.

certainty that a specific prescription will achieve its purpose. It is plausible that law should focus its data collection energy in areas where it presently uses standards, with an eye on converting them to rules. Both structured and unstructured inquiries can generate new, testable theories. When this iterative process leads nowhere, law can continue to use its loose standards and unrefined rules. In many cases, however, the study of good data will lead to a new standard, a tightening of a rule, or a switch from a standard to a rule. It is easy to imagine shifts from long-held practices. Thus, the Uniform Commercial Code's obligation that every contract or duty be performed in good faith, may fall to a tighter standard, such as parties cannot form contracts which they should know they cannot fulfill,[56] or even to a rule, such as one freeing breachers from liability (and relying on reputational interests) for contracts of more than $10,000. This sort of development could come about through the common-law process, but it seems more likely to be brought about with the deployment of AI. Topic modeling[57] is just one technique for evaluating unstructured data. By analyzing word counts and clustering across a collection of judicial opinions, a topic model can reveal specific factual instances where a standard such as good faith is repeatedly applied, and that revelation can lead to greater specificity in law.[58] It is unlikely that the conventional common-law method would use this information to alter legal rules. The larger point here is that AI is likely to cause some standards to give way to rules, and surely likely to refine inherited rules.

There are at least two ways to enhance data collection and their usage in law. Law can simply collect more data. But law can go further, and mandate that data be structured in order to facilitate its collection. For instance, law might require that driverless cars be placed on a network so that tortious behavior can be easily defined in relation to aggregated patterns of

[56] *See* U.C.C. § 1-304 ("Every contract or duty within the Uniform Commercial Code imposes an obligation of good faith in its performance and enforcement.").

[57] Topic modeling is an application of unsupervised learning that finds groups of items when the analyst is unsure or disinclined to specify a target of inquiry. JULIA SILGE & DAVID ROBINSON, TEXT MINING WITH R 89 (2017). The algorithm examines documents—such as judicial opinions, statutes, and regulations—and fits a topic model to the documents. For instance, a collection of opinions that have been identified as related to successor liability through a keyword search can be examined by the algorithm, which then classifies them into broad categories such as cases involving liability evasion, risk allocation, and bankruptcy. *See* Frank Fagan, *Successor Liability from the Perspective of Big Data*, 9 VA. L. & BUS. REV. 391, 391 (2015). Note that the cases are sorted by the algorithm and not coded by the researcher. If the latter were true, the analysis would be supervised, but important topics might be missed.

[58] *See* Fagan, *supra* note 34 at 26 (noting that legal doctrine can become more streamlined through algorithmic identification of judicial rationale).

automated driving.[59] Similarly, law might select a subset of consumers and ask for information about their preferences for various contractual terms in order to develop personalized default rules for a larger population.[60] As these examples suggest, more data can lead to greater specificity in rulemaking. These are examples of AI's leading to a higher ratio of rules to standards.[61]

As machines continue to develop the tools of causal reasoning, they will face other challenges, perhaps related to omitted and confounding variables (and certainly related to the identification of relationship structure when the effects of an intervention cannot be computed) that bring on the problem of reversals.[62] On the other hand, big data and the arrival of more precise models may virtually eliminate reversals in relatively stationary legal domains. In those cases, Human-AI combinations may warrant less overruling. When machine learning suffers from overfit and other future challenges because of indiscernible patterns in data and variation in context, then the balance may favor greater overruling. Inasmuch as so much of law is about disparate conditions and social change, it is easy to see how standards, that empower human decisionmakers, may have more of a future than optimists about AI imagine.

C. Rules to Standards

The migration just described can run in the other direction. Data can show more indeterminacy than previously imagined, and this can cause law to devolve from specificity to generality. By indeterminate we mean that the

---

[59] *See* Mark A. Geistfeld, *A Roadmap for Autonomous Vehicles: State Tort Liability, Automobile Insurance, and Federal Safety Regulation*, 1611 CAL. L. REV. 1611, 1612 (2017) (advocating that manufacturer liability for autonomous vehicles should be assessed according to aggregate fleet behavior).

[60] *See* Ariel Porat & Lior Jacob Strahilevitz, *Personalizing Default Rules with Disclosure and Big Data*, 112 MICH. L. REV. 1417, 1450 (2014).

[61] It seems apparent that greater magnitudes of processed and usable data will lead to greater rule specificity. For example, knowledge of traffic patterns can teach us where right turn on red is reasonable and where not. In this case an existing rule (rather than a standard, like "turn whenever it appears safe to do so") might be refined by big data. In the private sector with which we are most familiar, the measurement of longer-term student and societal outcomes can suggest specific rules for admitting applicants to a law school, just as more data on productive outputs might reveal the ideal gender composition of corporate boards.

[62] *See* Ilya Shipster & Judea Pearl, *Complete Identification Methods for the Causal Hierarchy*, 9 J. MACHINE LEARNING & RESEARCH 1941, 1941 (2008) (describing causal queries that cannot be computed from the observation of lower-level associative relationships, cause-effect relationships, or counterfactuals).

data reveals no correlation, or none that is replicable, where one was previously assumed. More interesting is the suggestion that the examination of data will reverse our thinking about the impact of law on behavior or outcomes.[63] A familiar rule may lose its empirical support, and abrogation or devolution toward a general standard may better serve its agreed-upon purpose. For example, a rule fashioned to increase diversity within a law school's incoming class, or a workplace, may simply prove wrong. Data may reveal unintended consequences of well-intended legal or private rules. In such a case, a more precise rule—or, counterintuitively, a switch to a standard—may be warranted. This pattern of lawmaking can be seen in the Federal Sentencing Guidelines. After an unsuccessful attempt at reducing racial disparity in sentencing decisions with well-defined rules, the law was changed to return some flexibility to judges.[64] This might prove to accomplish the intended result even if judges cannot explain the basis for their decision-making, and even where data analysis by the machine is unable to reveal the basis for these superior decisions – perhaps because these are not the same for different judges. It is tempting to say that the use of rules is categorically incorrect in the sentencing domain, and that standards are simply more appropriate. But this reaction is premature. As indeterminacy wanes with enhanced data collection and organization, the availability of precise rules that carry out their intended purpose will increase. The important caveat that we have emphasized in this Article is that a legal domain must remain sufficiently stable, or its dynamism must be at least identifiable. Inasmuch as law offers indiscernible patterns and changing contexts, machine learning cannot occur.[65]

For rules to overtake standards, either good theory (and then structured data) or good data is required.[66] Machines are likely to need humans or, put differently, the combination may outperform either human or machine acting alone.

V. CONCLUSION

Whatever aims one ascribes to, or wishes for, law, data can play a significant role is assessing the effectiveness of a given legal rule, its

[63] *See supra* notes 8-9 and accompanying text.

[64] *See* Nancy Gertner, *A Short History of American Sentencing: Too Little Law, Too Much Law, Or Just Right*, 100 J. CRIM. L. & CRIMINOLOGY 691, 698-707 (2010) (providing an overview of the history of the Guidelines and documenting the change from their mandatory to advisory role).

[65] *See supra* notes 21-22 and accompanying text.

[66] *See supra* note 2 (noting the importance of prior knowledge (or theory) in structured representations of the world).

enforcement, or even the performance of a legal system as a whole. Artificial intelligence can help lawmakers, and the citizens who evaluate them, exploit the available data to assess the impact of familiar and imagined legal rules on such things as wealth distribution, crime rates, employment rates, and national income. It will soon help us evaluate the impact of even modest changes in intellectual property law, tax law, and environmental regulations. But this Article has not sought to convince skeptical readers that better analysis of data will help lawmakers understand the effects of the rules they develop. It has focused instead on the impact of AI on the architecture of law. It has suggested that rules are likely to become more detailed, and that there will be a dramatic change in the ratio of rules to standards. On occasion, artificial intelligence will reveal that standards, administered by conventional regulators and judges, are actually superior to rules, designed by courts and legislators, aided perhaps by their machines, who will be anything but artificial.

These changes will rely on the collection of data; the development of AI suggests that we pay more attention to the scope and accuracy of this collection. As important, it is time for serious thinking about the instructions, or goals, communicated to artificial intelligence. We are likely to pay more attention to what it is, exactly, that we want law to accomplish, because the details of legal rules will be formulated by artificial intelligence as much as by humans. This is so even though artificial intelligence is inevitably imperfect, as we have explained in this Article. But the reasons for AI's fallibility, including overfitting, reversal paradoxes, and flawed or troubling goals (normally introduced by humans) are hardly new, inasmuch as human lawmakers can be criticized for making the same mistakes. The question for the future is not whether machines are perfect, but rather whether we can identify the tasks in which machines, combined with humans or acting alone after receiving instructions, are superior to the humans we currently glorify as judges, regulators, legislators, and enforcers.

* * *