MIT Computational Law Report

The Un-Modeled World: Law and the Limits of Machine Learning

Frank Fagan

Published on: Sep 06, 2022

URL: https://law.mit.edu/pub/the-un-modeled-world

ABSTRACT

There is today a pervasive concern that humans will not be able to keep up with accelerating technological process in law and will become objects of sheer manipulation. For those who believe that human objectification is on the horizon, they offer solutions that require humans to take control, mostly by means of self-awareness and development of will. Among others, these strategies are present in Heidegger, Marcuse, and Habermas as presently discussed. But these solutions are not the only way. Technology itself offers a solution on its own terms. Machines can only learn if they can observe patterns, and those patterns must occur in sufficiently stable environments. Without detectable regularities and contextual invariance, machines remain prone to error. Yet humans innovate and things change. This means that innovation operates as a self-corrective —a built-in feature that limits the ability of technology to fully objectify human life and law error-free. Fears of complete technological ascendance in law and elsewhere are therefore exaggerated, though interesting intermediate states are likely to obtain. For computational law, predictive accuracy is limited by the relative dynamism of a given legal domain. Computational law will ascend easily in closed legal domains, but will require continual adaptation and updating across those legal domains where human innovation and openness prevail.

Introduction

As did Jürgen Habermas, in his essay on *Technology and Science as 'Ideology'*,[1] I am proposing an explanatory scheme that, in the format of an essay, can be introduced, but not seriously validated with respect to its efficacy and merit. The oversimplifications of the problem of technological ascendancy and my reduction of various responses and solutions serve to clarify this scheme. They are no substitute for its scientific validity. Nevertheless it has become clear to me, albeit intuitively, that when thinking about the rise of machine learning and rationalized society generally, many legal scholars mistakenly presuppose the unlimited expansionary power of a machine's capacity to learn. It is as if one observes the moon landing, and then immediately anticipates full colonization of the universe on the basis of a "natural" and "ineluctable" trajectory of human achievement without ever considering the physical limitations of time and space.

It is true that there is a lively and ongoing debate over the technical feasibility of creating a machine that replicates the human mind (Tegmark 2017), but it is a deep irony that replication of our own intelligence is considered a significant technological ambition. Humans cannot perfectly predict the future. A machine copy of our faculties merely reconstructs this limitation, and the common solution is to inflate and multiply it. Perhaps a synergy could take hold of larger and more numerous artificial minds, but in most of the imagined cases, amplified "super-intelligence" amounts to nothing more than an enhancement (Bostrom 2014). Super-intelligence promises faster calculations, fewer errors, and more wide-ranging expertise—a super-charged version of today's skills—but without paradigmatic transformation. Nothing suggests that perfect prediction and radical learning are on the horizon.

Engineers may eventually give us the capacity to compute the entirety of the observed past, but there is no side-stepping endless change and the emergence of entirely new patterns of behavior. We may be able to compute the best moves in games of *go* and chess, but those moves are computed in closed environments with fixed rules. If Ke Jie changes the rules of *go*, so that pieces can be captured only at the edges of the board, he will easily defeat AlphaGo until the algorithm has had a chance to retrain. Legal environments evolve in unexpected ways—such as how Ke Jie might imagine and then implement a more fluid and dynamic set of moves for the play of *go*. Entrepreneurial and human innovation is built on such principals. We reward unexpected discoveries.

Leslie Valiant, winner of the Turing Award in 2010, has famously shown that machine learning occurs when two conditions are satisfied: learnable regularity and contextual invariance (Valiant 2013). Learnable regularity means that a machine can observe a pattern, perhaps when data is big enough or a model is expertly specified. Once a pattern is observed, learning still requires something more. The pattern must repeat itself in a sufficiently similar context. If circumstances change, then learning cannot continue. Contemporary thinkers must deal with Valiant's conditions.

Adding additional variables, so that all possible future states may be computed—including fantastic new environments and eccentric low-probability changes to current environments—is not a solution because input must precede output in order for computation to take place. Truly spontaneous action has no input. In the physical world, we know that there are limits to predicting the value of a quantity prior to its measurement (Heisenberg 1927). This is true even with a complete set of initial conditions.[2] Another way of thinking of the uncertainty principle is that there is no input that generates the true value. This means that truly spontaneous action exists; and that "change is far more radical than we suppose" (Bergson 1984 [1911]: 1).

Running sufficient experiments is not possible either because spontaneity upsets an understanding of what constitutes sufficient. In some lucky cases, the machine may be able to discover additional variables and then develop imaginative experiments on its own. In others, it will lack the prerequisite knowledge and imagination of the future, and some spontaneous action will slip through its net. In short, unplanned innovations guarantee that perfect predictability in law and elsewhere cannot be achieved.

It is true that dogmatizing upon "radical unknowns," or endeavoring to limit reason with reasoning, is similar to —as Hegel colorfully put it—trying to swim without venturing into the water, but what is needed today is far less daunting. Acknowledging the limits of machine learning nurtures an understanding of when and why incorrect, and especially antiquated, models will generate errors. There is an additional advantage. When computational legal scholars and technologically minded lawmakers acknowledge the un-modeled world and its limitations, we abate growing fears of an ascendant technological and empirical character of law. Persistent unknowns imply a constitutional legitimization of law by means of human intersubjectivity or other non-empirically based rationales. This acknowledgement, of course, does not imply a reduction in model building

just as the physical limitations of space do not imply the end of vehicles. Machine learning is a tool like any other, and should be embraced to the extent that the effort is worthwhile.

While this Essay is concerned with inherent limitations to prediction, there is overlap with the current debate on the possibility of deep and dangerous AI misalignments. For instance, Eliezer Yudkowsky and Richard Ngo recently discussed the possibility of catastrophic malfunction of a "super-intelligent" AI.[3] Both seem concerned with the architecture of reinforcement learning. Catastrophe arrives because an AI will stubbornly seek rewards. If AI is trained and rewarded for achieving predictive accuracy, however, "full" rewards cannot be achieved because perfect prediction remains out of reach. To be sure, misalignments can occur inasmuch as humans lose control of a less-than-fully rewarded AI, but the point is that we should be more concerned with the architecture and objects of reward than with the computational capabilities of the AI. If, for instance, an intelligence were rewarded for making unexpected discoveries, then the danger of misalignment would likely be far greater (for both AI inside the box and for ones that escape it).

Further elaboration of these ideas related to misalignments is left to future work. For the moment, machine learning's incursion into law and the administration of human relationships is predicated on the computational predictive accuracy of human behavior. AI can, and very likely will, make deep incursions into legal domains that are more or less static and closed like chess and *go*. Open and fast-moving domains, on the other hand, will befuddle a technology that relies on detectable regularities and contextual invariance. The pervasive concern in law that unchecked ascendance of a technical rationality will degrade law's human and intersubjective character may to some extent be true, but many would suggest that this is desirable in legal domains that are relatively static and sufficiently predictable (see, e.g., Fagan and Levmore 2019). A problem arises when the same technology is deployed in open and dynamic environments where it is ill-suited (Fagan 2022). Blind transposition of tools that work well in static domains generates errors in dynamic domains. But tool mismatching presents a categorically different challenge than the eclipse of human input in law.

Fear of law's embrace of machine learning draws heavily upon modern philosophy and sociology. But the paradigmatic concerns of Heidegger, Marcuse, and Habermas, discussed below, reflect a common tendency to believe that machines will be able to do it all. Reservations and unease associated with technology, and the belief that machines have the capacity to efface life and law's human character, are less serious and foreboding when one considers that humans innovate and things change. Technology is temporal; humans are intra-temporal. Innovation and human dynamism therefore serves as a self-corrective to the full eclipse and importance of human input in law, the economy, and indeed all sociological subsystems that are in a state of progress and movement.

I. Various Statements of the Problem

A. Martin Heidegger

"But at the same time enframing, in a way characteristic of a destining, blocks poiēsis." – Martin Heidegger, *The Question Concerning Technology*[4]

Heidegger is concerned that our urge to direct or administer ("destine") technology towards an ordering, arrangement, and illumining of the totality of being denies us the ability to hear the call of a more primal truth (Heidegger 2008 [1953]: 318, 333). By believing and adhering to the perspective that the role of technology is to reveal truth and precision so that we may expertly organize nature, we lose sight of a deeper authenticity. Heidegger asserts that the outermost danger of this mistaken way of understanding technology is that humans come to see themselves as objects directed or administered by means of our very own framing and perception. It is as if we are so taken with our contemporary predetermination (Ge-stell) of technological process as a means of organization that we are blind to other perspectives of its character. This blindness, according to Heidegger, is what objectifies and imprisons us. Technology claims us instead of the other way around. As we envision and enframe technology as the "destining" or administration of truth revelation we become the very objects of its administration, which simultaneously "drives out every other possibility of revealing" and uncovering other truths (Ibid: 332).

Once he establishes that the problem is self-imposed ignorance, the solution comes easy to him. Humans must first ponder the rise of this perspective on technology—one aimed at revelation and organization—and how such a presupposition entraps us. Self-awareness of the problem is an essential component of its solution: humans are "the one[s] needed and used for the safekeeping of the essence of truth", always "holding before our eyes the extreme danger" of our proclivity to drive out deeper truth (Ibid: 338). Heidegger then tells us that we may retain deeper truth and authenticity by pondering the poetic in life and resurrecting a sense of mystery: "the more questioningly we ponder the essence of technology, the more mysterious the essence of art becomes." (Ibid: 341). Art and self-awareness can save us because they restrain us from directing or "destining" technical process under the influence of an attitude toward perpetual revealing and organizing of the totality of being. Self-awareness ties us to the mast so to speak, and if art may cover our eyes for a time, it paradoxically allows us to see other perspectives.

While Heidegger is surely correct that openness and a well-calibrated awareness can save us from engaging in technological error, he is surely wrong that human activity within our own technological framing (Ge-stell) "can never directly counter this danger." (Ibid: 339). He is mistaken to insist that humans must move to other institutions and enterprises such as art in order to fend off our proclivity to know everything, model the future, and automate ourselves into the abyss. A change in perspective is not necessary, and it may, just as well, not be enough. Humans have begun to delegate mystery-making to machines. Already black-box algorithms befuddle lawmakers with puzzling and troubling suggestions, and AI artists dazzle and captivate collectors (Bogost 2019). Our perspective on technology as a process that reveals and organizes the totality of being carries with it the seeds of technological colonization of mystery. Heidegger can encourage us to set boundaries in order to

reach deeper truth, but we can just as easily choose to ignore him with machine-made blackboxes and art. Valiant's two conditions for machine learning shows that technology itself erects the outermost boundaries in response to nothing more than human innovation and contextual change. Heidegger is, therefore, on much firmer ground when he links our relationship to machines (or technical tools generally) to the never-ending flow of time. Technological instrumentality is temporal, but life is intra-temporal.

B. Herbert Marcuse

"[S]cientific rationality, translated into political power, appears to be the decisive factor in the development of historical alternatives. The question then arises: does this power tend toward its own negation—that is, toward the promotion of the 'art of life'?" – Herbert Marcuse, *One-Dimensional Man: Studies in the Ideology of Advanced Industrial Society*[5]

Marcuse's concern is with human flourishing. He thinks that technology carries within itself a priority for eliminating scarce resources and the need for chasing after the mundane. This priority, as with the early philosophers, is understood to be elevated over all others. Because humans have embraced technology for reaching this ideological goal, Marcuse reasons that we have claimed technology (or tacitly allowed it to claim us, it makes no difference) and that we will demand, of our own will, that it totalize law and life, at least until scarcity is eliminated and life becomes more luxurious and interesting. There is no need to attempt to tie ourselves to the mast, recognize mystery, and embrace art. Tying our hands is not possible so long as we must engage in utilitarian labor.

Marcuse insists that both technology's goals and procedures are ideological. The elimination of scarcity is an ideology, but so is the science of measurement, because, no matter how hard we try, measurement remains imprecise. Exact and truly objective measure is a physical impossibility. He does not suggest that modern physics denies or questions the reality of the external world, but that "in one way or another, it suspends judgment on what reality itself may be, or considers the question meaningless and unanswerable." (Marcuse 2002 [1964]: 155). The theory of relativity and similar methods of thought help us organize our thinking on reality and being, but they do not objectify it rigidly. These physical limitations of measurement place humans (and their drive toward abundance) at the center of all things: our idealization of objectivity and our methods of science demand that we define physical matter in terms of its possible reactions to human experiments. He says: "No matter how one defines truth and objectivity, they remain related to the human agents of theory and practice, and to their ability to comprehend and change their world." (Ibid: 170). Echoes of Heidegger are heard here.

Although humans remain at the center of the technological project, they wield a supposedly neutral and non-human science. Modern science prejudicially sees and shapes the world in purposive and practical terms as a result of its own internal logic. Technology is, at its core, a social project of purpose-fulfillment. The purpose is

the domination of nature and the improvement of living conditions. This purpose remains linked to our own subordination:

> Nature, scientifically comprehended and mastered, reappears in the technical apparatus of production and destruction which sustains and improves the life of the individuals while subordinating them to the masters of the apparatus. Thus the rational hierarchy merges with the social one. (Ibid.)

In technology, Marcuse rediscovers the voice of Jamestown: "If any would not work, neither should he eat." Technical rationality determines the relative value of individuals to the group. Administration and rigid empiricism replace the invisible hand and its looser and more liberal rules grounded in human relationships. Technical rationality is comfortable with limiting discretion for the granting of meals to the idle, even if exceptions to the rule are deserved. After all, assessing the merits of deviations takes time and computational resources.

Marcuse is interested in understanding how technology will end. Unlike Heidegger, he is skeptical of art's ability to elevate distinctly human values over technical ones. Art is absorbed into technological society (Ibid: 67-69). His appeal to language and critical thought is an updated form of Heidegger's admonition to become self-aware. But instead of understanding technology as a tool for revealing the totality of being and preserving mystery with art, Marcuse suggests that we must come to see technology as an instrument of destructive politics. He does not call us to action, however, and instead tells us that scientific rationality will extinguish itself as a social project of worth once humans no longer need to perform "socially necessary but individually repressive labor." (Ibid: 235). Freed by technology from ensuring our comfort and survival, humans will develop a new scientific *technique*—one without trans-utilitarian ends and the attendant effort toward solving for scarcity—a task that generates so much of our political strife.

Given environmental variation brought on by human intervention and circumstantial change, the automated labor that frees us must occur in sufficiently closed environments to complete the technological project as imagined by Marcuse. He does not consider that machines will perform less satisfactory for those tasks that take place in dynamic and fast-changing environments. Even supposing that abundance is achieved, there is no reason to think that individual and environmental dynamism will fade. The satisfied and wealthy continue to work and compete for one another's affections (Smith 1976 [1759]), and the environment continues to evolve.

Marcuse might say that machine learning can muddle through with sufficiently few errors generated by environmental variation. And at least in some areas of law and life, the change will be slow enough. Machines will be able to adapt, perhaps with or without the help of humans, and continue to ascend. But this is not his point. Technology will ascend, regardless of its technical mastery of nature and prediction, because our political orientation toward abundance will demand it. But myths must be believed. Marcuse too quickly casts aside the physical limits of objective observation. He is not concerned with technological mastery on its own terms. He simply shows that humans will believe in anything, whether divine or homemade, so long as it

provides abundance. On Marcuse's own terms, technology's ascent is therefore limited by its ability to deliver on that promise. If failures accumulate, belief will erode. A golden calf preceded the Ten Commandments, but even Commandments may come and go.[6]

C. Jürgen Habermas

"The process of development of the productive forces can be a potential for liberation if, and only if, it does not replace rationalization on another level. Rationalization at the level of the institutional framework *can occur only in the medium of symbolic interaction itself, that is, through* removing restrictions on communication*."* – Jürgen Habermas, *Technology and Science as 'Ideology'*[7]

Habermas, too, falls into the trap of seeing technology as a means to an end. For Marcuse, technology is a means to abundance and freedom from labor. For Habermas, it is a means for legitimizing political power. Long ago, societies built their institutional "superstructures" with consensual norms, whose meanings were clarified by ordinary language. This superstructure was continually reinforced linguistically through intersubjective dialog, mediation, and exchange of ideas—what Habermas terms "symbolically mediated action." For a while, the sociological subsystems that descended from this superstructure—importantly the economy and state—were also built and managed with ordinary language among humans. Over time, legitimization through language and "talking things through" came to be replaced with empirical rationales. Slowly, technology began displacing humans as the legitimizers of economic and state action. Production processes, for instance, could be justified on the bases of wealth maximization, environmental damage minimization, and equitable income distribution. State action, like economic action, came to be rationalized along the same technical bases. For a while, technical and empirical rationales remained confined to the economy and state apparatus, while overall direction of society continued to emanate from people, and in particular, their intersubjective rationalizations of the institutional framework that encapsulates the economy, state, and other sociological subsystems. Today, Habermas tells us, the situation is reversed. Superstructure is no longer determined by deliberation among humans. It, too, is empirically rationalized.

It appears that this reversal came about primarily through advances in capitalism. Capitalism, notes Habermas, led to the collapse of the ideology of just exchange between buyers and sellers. No longer legitimized by free contract between two people engaged in deliberation and dialog, exchange must now be legitimized by political control and state regulation. This work puts the state in the uncomfortable position of directing economic stability and eliminating market dysfunction. Old-style politics defined itself in relation to practical goals; today's politics is defined in terms of solutions to technical problems. People can, and often do, stand in the way, which generates a need for the de-politicization of mass populations. Habermas contends that this is accomplished by replacing the bourgeois ideology of free exchange with an ideology of scientific and technological legitimization (Habermas 1971: 105). Technology, as the great legitimizer, is a means for justifying political decisions. In the process, it displaces communicative action and human intersubjectivity as means for the same.

Habermas wants us to protect intersubjectivity—with all of its untidy deliberation, dialog, and debate—from empirical overrun. A world in which language has been absorbed by technology is a world denatured. He shares this concern with Heidegger who saw in technology the eclipse of mystery. But instead of turning to art and its longing for deeper truth, Habermas sets his sights directly on the liberation of language:

> Public, unrestricted discussion, free from domination, of the suitability and desirability of action-orienting principles and norms in the light of the socio-cultural repercussions of developing subsystems of purposive-rational action—such communication at all levels of political and repoliticized decision-making processes is the only medium in which anything like "rationalization" is possible. (Habermas 1971: 118-19)

He sets aside Marcuse's solution of abundance on the grounds that we cannot deliberate societal goals, even with bellies full and in satisfied comfort, while we remain beholden to a restricted form of communication. Life is not worth living depoliticized. The words of Socrates on trial still haunt us—even if mangled.

But preservation of a liberated form of language is not the work of tending a garden. One cannot separate intersubjective and empirical rationales so cleanly. While Habermas is surely correct that empiricism is in ascendance, intersubjectivity is preserved through creativity and the expansion of novelty. Protection of language is less important in areas of law and life that are closed and predictable, areas where instrumental action will easily coopt and absorb deliberation and dialog. At the moment, this process appears engulfing because it, too, is novel. Once predictive learners are widely deployed throughout the sufficiently static domains of human activity, machine learning's ascent will abate. The same is true for computational law. Today's areas of human society and law that are open, changing, engaged in creation, and characterized by novelty remain—by pragmatic need—intersubjective. As those domains settle and begin to close, machines will provide their comparative advantages, and humans will move yet again to a novel activity at a further point in time. Technological rationality is not a means after all. It is a process of provocation because it is temporal and we are not. Even if we create a dynamic technology (like a dynamic learning machine), its ability to adapt is anchored to the moment of its creation, always asking us if we wish, or are able, to change it.

II. The Un-Modeled World

Understanding how humans interact with law allows a society to select and change its rules in accordance with its goals.[8] Some rules are more effective for achieving a goal than others because human behavior responds to incentives. Simpler rules for drafting a will, for example, may help a parent bestow assets more easily without the assistance of a lawyer. Slightly more complex rules may reduce disputes among children, but require expert advice. Overly complex rules may discourage the drafting of wills altogether, and so on. In order to select among rules so as to reach social aims, lawmakers deploy theories and models of human behavior. Models of the world help us predict responses to a change in law, which allows us to select the policy most likely aligned

with society's goals. It should be immediately obvious that if humans change, then the law must change, too. Machines need to keep pace.

Machine prediction may be imperfect, but if a machine can do the job better than a human, then it may be wise to cede some lawmaking control to machines.[9] Accurate prediction with machines requires observable regularities and stable contexts (Valiant 2013). Without both, learning cannot occur and prediction fails. Valiant describes the need for observable regularities as follows: "It requires that some regularity exists, and that this regularity be effectively detectable for any example." (Ibid: 62). Valiant describes his second condition for learning, the need for contextual stability, as the requirement "that predictions hold for examples drawn from the same source as the examples were drawn during learning." (Ibid.) If, for instance, the patterned behavior in one's home jurisdiction is different in another, then the experience of administering law at home teaches little for its administration abroad. The same limitations over space exist over time.

Valiant's two assumptions imply three hard boundaries for the use of machine learning in law: (1) inherently small sample sizes, which limit the ability to observe a pattern; (2) the reflexive behavior of humans and a tendency toward outsmarting rules, which accomplishes the same; and (3) our contemporary state of perpetual innovation and dynamism, which breaks down both patterns and contexts, especially over time. In legal domains where these three boundaries are trivial, machine learning and computational law will ascend easily. Otherwise, humans will remain central.

A. Small Sample Sizes

To the uninitiated, the idea of small sample sizes can be misleading. Lawyers, lawmakers, and legal scholars are often, unfortunately, in possession of a basic understanding of computer science and statistics. We think that small samples are a hindrance that can be readily overcome by collecting more data, and that additional collection is carried out easily, if not swiftly, so long as one knows where to look. It is true that if a statistical task proves difficult, it is usually because one does not possess a good source of data and does not know where to search. But this creates the illusion that *all* challenging statistical problems arise because data is hidden from us. If success eventually comes, and what is sought is found, we further stimulate a false belief that all necessary data can be uncovered by devoting and intensifying our efforts. After all, what is modest in life is often temporary. With personal commitment and the help of friends, the small can be augmented and raised to great heights. Surely the same can be accomplished for data.

This is not true. There are entire classes of data, required for prediction, that do not exist and never will. Consider "unlabeled outcomes." This term can also be misleading, because it readily implies a solution: one can observe the outcomes and label them. But observation is not possible because these "unlabeled" outcomes do not occur. They are counterfactual and hypothetical events that must be probabilistically assessed. Most often, they are merely the objects of prediction. Algorithms predict the recidivism of the prisoner who is

hypothetically released or the confusion of a trademark if it were permitted to circulate. Still, we never observe what would have occurred had the incarcerated criminal or enjoined trademark been liberated.

Contemporary methods improve on these and similar algorithms by comparing differences in labeling between judges. An algorithm that assesses a person or trademark as high-risk for recidivism or confusion suggests incarceration and injunction. But a lenient judge may decide to release. Data scientists can then observe the high-risk releases, which permits observation of what the algorithm suggests should remain unobserved. The process may be compared to a risky experiment. And it partially side-steps the problem of unobserved counterfactuals: we can see how some (but not all) high-risk defendants behave when released. We can never, unfortunately, observe the behavior of the others, classified as high-risk by a stricter judge, because their behavior upon release will always remain theorized and hypothetical. The recidivism label of the never-paroled inmate is indeed an *unlabeled* outcome, as is the confusion label of the never-circulated trademark, but an improved and more accurate term might convey the *impossibility* of empirical observation. Unlabeled outcomes are not merely missing their labels; they are missing the underlying events in the lived world that could generate those labels.

All methods for resolving this obstacle (if it can be thought of as one) rely on selecting features of the observed outcomes, and then matching those to the unobserved event in order to *impute* to it a possible outcome. This method works well if we understand the composition and importance of an outcome's features. But if we select the wrong features to observe, or fail to match those features correctly to an object of prediction, then the accuracy of algorithmic judgments will fall. "Empirically minded legal academics are often impressed with large data sets when it is the quality of hypotheses that matters most." (Levmore and Fagan 2021: 408). This is easy to see with an outlandish scenario. If a group of prisoners that possess a uniquely shared feature are never released, then a category of unobserved counterfactuals will persist, which can generate errors inasmuch as the underlying hypothesis supporting the group's continued incarceration is wrong. In a world where judges incarcerate all criminal defendants of above-average height, for instance, we cannot observe the post-release behavior of tall parolees. We are selecting the wrong features for matching.

Suppose that we select the correct features. For some legal questions, sample sizes are truly small, so much so that a pattern cannot be detected and machines cannot learn. State-of-the-art machine learning requires 5,000 labeled examples per category in order to approximate human learning capability (Goodfellow et al. 2016). Millions are needed to exceed it. With anything less, data scientists must rely on hypotheses and assumptions even more so. Consider that of all of the prisoners released in Wisconsin in 2011, only 2,379 proceeded to recidivate over the next three years.[10] One can hypothesize similarities over time and space in order to increase their number so long as those released in 2011 are comparable to those released in say, 2021, or defendants in Wisconsin behave sufficiently the same as defendants in say, Florida. Of course these comparisons are useful because the relationship between the features of recidivists to their behavior is sufficiently understood. We have selected the correct features for matching. But the point is that detection of regularities can prove elusive

when domains are under-theorized. Persistently small sample sizes clearly obstruct the detection of patterns, but so does poor theory in the face of mountains of evidence. We must comprehend what can and cannot be compared.

B. Reflexive Behavior

Humans act in anticipation of the future. Anticipatory and interdependent action distorts learning because people modulate their behavior in response to others. What is observed is partly based upon an intervention but also upon the reaction to it, or its mere announcement. If Wisconsin builds a dam to stop flooding from a local reservoir, we may think that we are protecting a community of 2,000 families. But we must estimate how many additional families may move to the newly protected community. The number may grow to 2,200. We must also consider that current residents may move away, say for instance, if the dam is unattractive and they prefer beauty. Building the dam for the current residents will only protect 2,000 families if the behavior of future and current residents is unaffected by the dam. Previous observations are unhelpful when alternative rules, policies, and legal environments are introduced. They must be recalibrated, on the basis of a model of the world, in order to retain their predictive value (Lucas 1976). Machine learning and its applications in law must reckon with reflexive behavior.

Social scientists, particularly economists, create “structural models” that explicitly represent a person’s preferences, or relative desires, and specify how these are shaped and formed. As statistical representations of human taste formation, these models are reductive, but they can tell us something useful about how a person will react to an intervention. In the past decade, machine learning has been instrumental for estimating structural models because data can help us partly understand how people form tastes in response to programmatic interventions. But it bears repeating that in order for any prediction to achieve accuracy, the model must detect a pattern within a sufficiently stable context. With sufficient data and expert modeling of preference formation, machines are able to detect patterns of human response to programmed interventions, but machines remain susceptible to error when unimagined change disrupts human desire and its formation.

This remains true whether machines deploy predictive or causal reasoning. A machine may learn that imposing the construction of larger prison yards causes a reduction in prison violence, but if a pandemic arrives and prisoners begin practicing social distancing on their own, increasing the size of prison yards may no longer cause reductions in violence, and certainly not to the same extent prior to the pandemic. Inasmuch as the environment is a relevant input to a structural model, its causal conclusions depend on sufficient contextual invariance just as well.

C. Innovation

By now it should be clear that machine learning cannot predict the future. “All it can do is to map out the probability space as it appears at the present and which will be different tomorrow when one, of the infinity of possible states, will have materialized.” (Gabor 1963: 184). This map, even if finite, can surely be improved.

Small sample sizes can grow, and increasingly perceptive observers and modelers can better represent human behavior. On the map of possible outcomes, probabilities will sharpen, newly revealed outcomes will be added and obsolete ones discarded. But the map just as surely remains finite: some class of outcomes will remain out of reach inasmuch as humans innovate and things change.

Of course behavioral and environmental changes can be, and often are, interdependent. External shocks to circumstances provoke human adjustment and ingenuity. Even so, invention builds on that what preceded it. Perhaps it can be serviceably understood in terms of biological reproduction. By those same terms, we see that invention can mutate, spontaneously, in unexpected ways (Leroi-Gourhan 1945: 344). It is true that spontaneous interventions are overlain and combine with artifacts, but the underlying processes and objects cannot chart the trajectory of an artifact struck with spontaneity. As with quantum states, spontaneous arrivals must be observed before they are measured.

These types of unknown unknowns which populate the map of infinity can materialize when spontaneous changes to the environment provoke new and unexpected human behavior. They also appear when humans carry out experimental or random acts on their own. Every social and technological innovation broadens the probability space of future behavior. As we discover and learn, sometimes spontaneously, previous horizons contract and new ones expand. Facing infinity is the price of admission for moving forward in time.

III. Criticisms

A. The Infinite Model

The un-modeled world would disappear if we could map the full probability space of the future. With sufficient data and computational power, entire universes of possibilities could be estimated, and if any error were to arise, that too, would be fully predicted within an interval of confidence. A society, whether capitalist or socialist, could achieve precise allocation of its goods and services on any basis or goal that it selects. A legal system, of any constitutional order or tradition, could create and administer precise and well-tailored computational rules, which could be personal and contingent on any circumstance that may arise. If data is too small, new observations will accrue over time. If circumstances change, so that the data becomes stale, then models can be deployed to enliven it. This is an infinite model. It confronts small sample sizes with additional effort, models unobserved counterfactuals and reflexive behavior with skill, and supposes that it can see deeply into human desire. In its enthusiasm to map the entirety of human behavior and its context, it bypasses spontaneity and relegates it to error.

B. The Closed Model

The “errors” of the un-modeled world become unimportant when one actively chooses to close the full probability space. By setting boundaries on the map, learning is focused on what is seen. Machines placed within fixed environments are able to sharpen their predictive precision. Strategies of closure track the earlier

examples of games. The rules of play and their context (the game board) are fixed. Because innovative adjustments to player behavior and the game board are bounded, a machine can easily compute the entire universe of possible moves. It can distill optimal and contingent policies in relation to its opponent's moves as Zermelo theorized a century ago.

Competent data scientists limit the probability space as a matter of course. Consider, for instance, a model of 10 Boolean variables. The number of possible combinations is 2^{10}, or 1,024. Increasing the number of variables by just one, doubles the number of possible combinations to 2^{11}, or 2,048. The COMPAS algorithm predicts recidivism on the basis of 137 variables. Not all are Boolean, but assume that they are in order to begin to understand the magnitude of the number of combinations (or hypothesized states of prisoner features that can most accurately predict behavior): 2^{137} yields 1.74e+41. When the algorithm considers an additional variable, the number of possibilities doubles to 3.48e+41. Potential models of recidivism grow exponentially with the addition of a single variable (even if it is merely Boolean), and observing a new defendant often introduces new variables inasmuch as humans are unique and present different features. To manage the combinatorial explosion, data scientists deploy hypothesis testing to eliminate candidate models (Valiant 2013). Excluding the unlikely and setting boundaries to the map allows for learning and prediction. This is sensible when we possess a theory of behavior and are convinced that its context is relatively invariant. As our understanding grows, we can more confidently direct the model to ignore certain features of the probability space. Perhaps, after all, the known world is expanding and the unknown world is contracting. An interval on the real line can become infinitesimally small, even as it resists point-identification. Maybe that is where social spontaneity will persist. But technological and social advance broadens the probability space. Our own advancement expands the horizon of possibilities endlessly.

C. The Updating Model

Computational law and machine learning will chart a course nonetheless. Charting will be accomplished by keeping up with the dynamic and ever-changing probability space. Old locations on the map will submerge and disappear as patterned behavior and contextual change renders them obsolete. New locations of stability will appear, at first with small data, with humans at the center. As data accumulates and patterns are uncovered, decision-making will be delegated to machines in order to take advantage of their superior precision. Humans may work around the machines, but structural models and enhanced techniques that clarify human desire will provide further precision. Machine learning will remain stable and prediction will succeed for a while. As the probability space begins to change, errors will arise. Models will update in response by adding new variables, discarding others, and recalibrating their organization and shape. We will abide by Galileo's maxim to "measure what is measurable, and make measurable what is not so," but what is made measurable will be made un-measurable again and again. Perhaps models will update on their own—powered by sensors, antennae, and other automated measuring tools—but these, too, will become obsolete in unimagined ways.

How much of the map can be closed will become an important question for data scientists and computational lawyers. Some periods may be more innovative than others if the past is any indication. Periods of renaissance and stagnation tend to oscillate. Perhaps the updating model can adapt quickly as human populations swell and surveillance becomes accepted or imposed. To the untrained eye, the updating model may even appear infinite. How much of the map should be closed is already an important question for law. Rules limit human discretion, even when accuracy demands it.

Conclusion

The predictability of all things is a mechanical delusion. Small sample sizes, reflexive human behavior, and innovation all limit machine learning and computational law's predictability. In some instances, these hurdles can be overcome. In others, they cannot. Patterns must be detectable and their contexts must remain sufficiently stable. But things change, and at times radically so. Machines are capable learners in closed environments, but they must struggle to keep up in fast-moving ones. Law and life will benefit from machine learning in sufficiently slow-changing domains; humans and their discretion will remain central when novelty is present.

Concerns that law is losing its intersubjective character are tied to concerns with technological ascendancy. Inasmuch as the rise of technology is tethered to machine learning, it faces inherent limitations. Philosophers like Heidegger encourage us to set boundaries to technological incursions by maintaining a sense of mystery through art, but Valiant's conditions for learning erects boundaries without our assistance. Political theorists like Marcuse suggest that the technological project will end once it solves for scarcity, but even if this is possible, humans continue to find ways to contend for one another's affections. Belief in technology will erode before it successfully yields any utopia. Sociologists like Habermas observe a process of colonization: technological and empirical rationality is driving out human intersubjectivity. He encourages us to respond by liberating our communication from technological rationales and then maintaining control of language. But machines, too, enter into dialog and modulate our communication. Intersubjective rationality is effortlessly maintained through innovation and spontaneous change. Its overreach is inherently limited.

Machines can only learn if they can observe patterns, and those patterns must occur in sufficiently stable environments. Without detectable regularities and contextual invariance, machines remain prone to error. Fears of complete technological ascendance in law and elsewhere are exaggerated, though interesting intermediate states are likely to obtain. The infinite model is a delusion because the probability space of life is expanding. Closing the model may provide precision for a time, but errors will eventually accumulate. Although the updating model can never fully map the un-modeled world, it is able to recalibrate to contextual variation. How quickly it can adapt is not a technological or engineering problem. Adaption speed varies with the rate of unexpected human innovation and contextual spontaneity. The un-modeled world cannot be overcome.

References

Bergson, Henri. 1984 [1911]. *Creative Evolution*. Arthur Mitchell (trans.) (Lanham: University Press of America).

Berman, Harold J. 1974. *The Interaction of Law and Religion*. (Nashville: Abingdon Press).

Bogost, Ian. 2019. "The AI-Art Gold Rush Is Here." *The Atlantic*.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. (New York: Oxford University Press).

Fagan, Frank. 2022. "Law's Computational Paradox." unpublished.

Fagan, Frank, and Levmore, S. 2019. "The Impact of Artificial Intelligence on Rules, Standards, and Judicial Discretion." *Southern California Law Review*, 93: 1-36.

Gabor, Dennis. 1963. *Inventing the Future*. (London: Secker and Warburg).

Goodfellow, Ian, Bengio, Y., and Courville, A. 2016. *Deep Learning*. (Cambridge: MIT Press).

Habermas, Jürgen. 1971. "Technology and Science as 'Ideology'." In Jeremy J. Shaprio (trans.), *Toward a Rational Society: Student Protest, Science and Politics* 81-122 (Boston: Beacon Press).

Heidegger, Martin. 2008 [1953]. "The Question Concerning Technology." In David Farrell Krell (ed.), *Martin Heidegger: Basic Writings* 311-41 (New York: HarperCollins Publishers).

Heisenberg, Werner. 1927. "Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik." *Zeitschrift für Physik*, 43: 172–198.

Leroi-Gourhan, André. 1945. *Milieu et techniques*. (Paris: Albin Michel).

Lucas, Robert J. 1976. "Econometric Policy Evaluation: A Critique." *Carnegie-Rochester Conference Series on Public Policy*, 1: 19-46.

Marcuse, Herbert. 2002 [1964]. *One-Dimensional Man: Studies in the Ideology of Advanced Industrial Society*. (New York: Routledge).

Prevedel, Robert, Hamel, D.R., Colbeck, R. et al. 2011. "Experimental Investigation of the Uncertainty Principle in the Presence of Quantum Memory and Its Application to Witnessing Entanglement." *Nature Physics* 7: 757–761.

Smith, A. 1976 [1759]. *The Theory of Moral Sentiments*. (New York: Oxford University Press).

Tegmark, Max. 2017. *Life 3.0: Being Human in the Age of Artificial Intelligence*. (New York: Vintage Books).

Levmore, Saul and Fagan, F. 2021. "Competing Algorithms for Law: Sentencing, Admissions, and Employment." *The University of Chicago Law Review* 88: 367-412.

Valiant, Leslie. 2013. *Probably Approximately Correct*. (New York: Basic Books).

Voltaire. 1776 [1736]. *L'Examen Important de Milord Bolingbroke*. (London: no imprint).

Footnotes

1. In Jeremy J. Shaprio (trans.), *Toward a Rational Society: Student Protest, Science and Politics* (Boston, MA: Beacon Press 1971). ↩

2. Uncertainty can be reduced with various techniques, but it has never been eliminated (see, e.g., Prevedel et al. 2011). ↩

3. See Scott Alexander, AstralCodexTen.substack.com, Practically-A-Book Review: Yudkowsky Contra Ngo On Agents (Jan. 19, 2022). ↩

4. In David Farrell Krell (ed.), *Martin Heidegger: Basic Writings* 311-41 (New York, NY: Harper Perennial Modern Thought Edition 2008). *The Question Concerning Technology* arose from a series of lectures given between December 1949 and June 1950. The lectures were revised into the essay in November 1953. ↩

5. Marcuse 2002 [1964]: 235. ↩

6. Indeed, Marcuse may be unsurprised if a resurrected Voltaire, who said that "[e]very sensible man, every honorable man, must hold the Christian faith in horror," were to eventually say the same of machine learning. (Voltaire 1776 [1736]: 203). ↩

7. Habermas (1971: 118). Emphasis in the original. ↩

8. Here I use law, rules, and policies in a broad sense to mean substantive rules and standards that govern behavior, procedural and administrative rules that govern the application of those substantive rules, constitutional rules that shape their interpretation, and so on. Similarly, I collapse the term goal to a social objective, which can include the balancing of multiple goals. ↩

9. Of course humans could cede the task of goal-setting to machines, in order, for instance, to place humans behind a veil of ignorance. The propensity to follow such rules, however, is tethered to human preferences since the possibility for revolt remains (Berman 1974). See also (Fagan 2022). ↩

10. Wisconsin is of interest because it was the location of an early legal challenge to that state's use of algorithmic risk-assessment. ↩