Chapter 6

Paths to Responsible AI: Reflections from the Classroom

Jeffrey W. Chivers, William N. Eskridge, Jr., and Theodore I. Rostow[1]

The primary conviction that informed our launch and design of Artificial Intelligence, the Legal Profession, and Procedure at Yale Law School was that advancements in artificial intelligence (AI) will substantially influence the careers of the next generation of lawyers, jurists, and legal academics. The rapid advancement in 2023 and 2024 in the scale and capabilities of large language models (LLMs) has underscored the need to broaden and deepen understanding of AI in both legal education and the profession. There is likewise a heightened need to develop practical frameworks to evaluate the concrete benefits and risks that accompany the use of AI in specific legal contexts. Like others, we are also interested in the medium- and long-term consequences of AI for the practice of law, the legal profession and its ethical requirements, and our system of adjudication.

I. Technical Basics of Generative AI

Lawyers and legal professionals do not need to become technical experts in generative AI to use it responsibly in legal studies and legal work. A basic understanding of how generative AI is different from other forms of software, however, is critically important to making informed decisions about when and how to use this technology. In the classroom, we have found that three central concepts are

1. William N. Eskridge, Jr. is the Alexander M. Bickel Professor of Public Law at Yale Law School. Jeffrey W. Chivers is the CEO of TLATech Inc., the managing partner of Chivers LLP, and a visiting lecturer in law at Yale Law School. Theodore I. Rostow is the COO at TLATech Inc., a partner at Chivers LLP, and an Irving S. Ribicoff Visiting Lecturer in Law at Yale Law School. Professor Eskridge and Messrs. Chivers and Rostow cotaught Artificial Intelligence, the Legal Profession, and Procedure in the spring 2023 and fall 2023 semesters at Yale Law School.

instrumental to help law students without a technical background to begin to think critically about AI in a manner that is grounded in the technology.

A. Traditional Programs vs. Machine Learning Programs

Traditional methods of software programming relied upon and tracked fairly closely the constraints of conditional, rules-based logic—that is, if a certain condition or criterion is satisfied or not satisfied, then the program performs certain operations or provides certain outputs. This paradigm of software programming underpins the earliest chatbots (such as "Eliza," the "therapist" chatbot from the 1960s) as well as more complex software systems in the legal domain, such as TurboTax. The traditional software paradigm—rules-based logic—underlies the vast majority of software that we have become accustomed to using over the last 40-plus years. Programs written in this manner are predictable in their behavior and capable of being audited and understood by humans with sufficient expertise.

The technique of machine learning, by contrast, produces software programs whose behavior is not governed strictly by rules-based logic written by human programmers. Unlike the traditional method of software programming, in machine learning the human programmers write logical rules and procedures for a program to follow in order to learn mathematical patterns from (usually large amounts of) data. The behavior of a program that derives from machine learning, such as a traditional predictive coding model (as in Technology Assisted Review, or TAR) or a large generative AI model, is governed by the mathematical functions that were learned by the program from its training data.

The software products available to law students and lawyers increasingly will rely on a combination of conditional, rules-based logic *and* behavior driven by the mathematical functions that were learned through machine learning. Developing an understanding of the distinction, and an awareness of the trade-offs, between these two forms of programming is an important step toward being able to evaluate responsible uses of generative AI.

B. Representation of Human Language in Vectors

The second concept that is central to bridging the technical divide among some of our students is the concept of *vectors* for the representation of human language.

Computers are fundamentally machines that perform math. Although many layers of abstraction now separate software programs (let alone human users of software) from the underlying hardware of computer systems, that hardware performs basic mathematical operations, such as addition, subtraction, multiplication, and division. For computer systems to exhibit understanding and sensible

generation of text in a natural language, such as English, they must convert natural language (the words, sentences, paragraphs, etc.) into a form of representation on which mathematical operations can be performed.

The form of representation that LLMs use to represent human language is called a *vector*, which is a (usually) long sequence of decimal-point numbers. For example, one language model trained on the entirety of Wikipedia represents the English word "cat" as the following 300-dimension word vector:

0.007398007903248072, 0.0029612560756504536,
–0.010482859797775745, . . . , 0.0001564373378641903420.[2]

The following question often arises for students interacting with this material for the first time: What does each of the numbers in these word vectors correspond to? The remarkable answer is that, in general, humans cannot identify what a given number in a word vector signifies, nor can we determine what the overall set of numbers in a word vector signifies exactly. We have some limited understanding of word vectors—we know, for example, that word vectors are trained so that similar words (words that can often be replaced by each other in a sentence or words that tend to occur in similar contexts) will have similar numeric representations (the vectors will be "close" in high-dimensional space)—but a more comprehensive understanding of the meaning of these vectors is, so far, elusive.

When LLMs are "reading," "analyzing," and "generating" words, they are interacting mathematically, in whole or in substantial part, with these vectors. Generally speaking, the first step in a language model's process for generating text is the language model taking input text (assuming that is the format of the input) and converting the text into vectors.

C. Neural Networks

The final concept we have found essential to bridging the technical divide is the concept of a neural network. The term "neural network" refers to a software architecture, modeled roughly on the human brain, that has become a centerpiece of modern machine learning systems. A "neuron" in this context refers to a single node within a neural network. Each node in the network performs the simple task of taking numerical inputs from other nodes, applying a mathematical function (such as cosine or tangent), and passing the numerical results on to one or more

2. The 300 decimal numbers that make up this language model's word vector for "cat" are available at http://vectors.nlpl.eu/explore/embeddings/en/MOD_enwiki_upos_skipgram_300_2_2021/cat_NOUN/ (click on "Show the raw vector of «cat» in model MOD_enwiki_upos_skipgram_300_2_2021").

other neurons in the network. Although each individual neuron does a simple calculation, the aggregation of many layers of neurons can yield neural networks that, when considered end to end, embody extremely elaborate mathematical functions. In the aggregate, these functions map numerical inputs to numerical outputs.

Descriptions like these often leave students asking how a complex mathematical function, which operates entirely on numerical inputs and numerical outputs, can result in software that can engage in conversational dialogue, convert prose into outlines, summarize the key points in a document, or demonstrate other verbal capabilities. The core mechanism that bridges this divide—the divide between human language, on the one hand, and the mathematical operations that comprise a neural network, on the other hand—is the clever transformation of natural language tasks (answering a question, engaging in dialogue, summarizing a document, etc.) into the comparatively smaller task of predicting the next word in a phrase, sentence, paragraph, or passage. To predict the next word in a segment of text, the neural network of a generative AI takes the inputs (natural language, represented as numerical vectors), applies the mathematical functions embodied in the neural network (which were learned from patterns in the neural network's training data), and makes a statistical prediction as to what word (or symbol, such as punctuation) is most likely to come next. In generative AI systems with a broader range of "skills," this core mechanism—that is, predicting the next word, and the next, and the next—is cleverly extended to perform other natural language tasks—such as summarization, classification, question answering, and so on—by transforming the desired natural language tasks into the more basic, and purely mathematical, problem of predicting the next word, and then the next, and then the next.

The core mechanism that underlies current generative AI systems—that is, predicting the next word in a passage of text—has prompted a substantial divide among researchers and practitioners of AI as to how intelligent modern generative AI systems really are. Due to the statistical nature of this core mechanism, generative AI systems have been criticized as being nothing more than "statistical parrots," which can merely mimic intelligence in some instances. According to the extreme version of this view, recent advancements in generative AI are mostly illusory, and the current enthusiasm around generative AI will, a few years from now, be understood as greatly overhyped. But the rejoinder to this extreme view is also straightforward—at what point, in learning accurately to predict the next word in billions of sentences, does the neural network develop a higher-order understanding of the concepts, relationships, hierarchies, and principles that conceptually drive the next word in a sentence? This open research question is one that we encourage students to understand and follow, as the type and degree of

"intelligence" embodied in the neural networks of generative AI (are they statistical parrots, or do they develop higher-order representations of the world that is described by the content of their training data?) will substantially influence, at the most basic level, how profound an impact generative AI will have on knowledge industries, including the legal industry.

In summary, generative AI outputs are governed by the operation of math functions, adjusted through a machine learning process, rather than conditional logic rules programmed by humans.[3] For many tasks and workflows in the legal context, software programs or software components that are governed by conditional logic rules are superior to software programs or components that utilize machine learning or generative AI: the outputs of conditional logic programs are transparent, predictable, and capable of being audited. For other tasks and workflows, a nuanced combination of conditional logic rules and generative AI yields the best results. Understanding the capabilities and shortcomings of generative AI begins, we respectfully submit, with an understanding of how it differs from conditional logic–oriented programming.

II. Generative AI Will Change the Legal Profession

In 2023 and 2024, there has been an explosion in the number of generative AI products being offered to the legal profession. The range of proposed uses of generative AI in legal work will no doubt expand further in coming years. Students and lawyers will benefit from conceptual frameworks for evaluating the intersection of generative AI with legal practice and the profession more broadly to make sense of the deluge in new tools (and the hype around them); to take advantage safely and ethically of generative AI capabilities in their work; and to help, in the aggregate, guide the trajectory of a legal profession increasingly empowered by generative AI.

A. Evaluation of AI Use Cases

Conceptual frameworks may aid lawyers, students, courts, and regulating bodies in their efforts to evaluate specific generative AI use cases (i.e., the use of generative AI to replace or support existing workflows). Without purporting to establish a single, one-size-fits-all framework for such evaluation, we suggest that frameworks designed to aid in the evaluation of generative AI use cases could take

3. Importantly, language models can be integrated into applications that also use traditional conditional programming techniques, including in chatbots, such that the "output" that the user sees is the result of the combination of language model operations and traditional, event-driven programming.

into account (1) the context in which generative AI would be used, (2) the task that would be performed by generative AI, (3) the purported gains in quality or efficiency of that task, (4) the types of errors that could be introduced by the use of generative AI, (5) the error (in)tolerance and duties of care that attach to the task at issue, (6) the costs of the generative AI approach, (7) the extent to which human validation must be performed, (8) whether the use of the tool in the context at issue would undermine procedural fairness or circumvent ethical requirements, and (9) the process changes (if any) that attend the use of generative AI.

In addition to weighing these considerations, academics and practitioners alike need to perform substantial work to evaluate empirically the performance of generative AI–powered systems in connection with legal tasks. This process means rigorously studying how different AI models or systems perform different legal tasks in different contexts.

There is substantial untapped potential for academics and practitioners to share resources and collaborate in evaluation efforts. The ABA and other organizations should promote these collaborations consistent with existing ethics requirements. Careful, rigorous study is an important countermeasure to "Legal AI" hype, which will eagerly present "Legal AI" as a solution to existing workflows (irrespective of whether the tools can actually serve as adequate substitutes or even adequate complements to existing workstreams). There is a history in legal tech of companies making bold claims about the performance of their software that are not independently verifiable (and, in time, are proven to be inaccurate). The profession should carefully evaluate claims about the capabilities of "Legal AI."

B. Broader Consequences for the Practice of Law

As students, lawyers, firms, and decision-making bodies begin to incorporate generative AI into their work, substantial consequences for the legal profession and society will result. Looking ahead to the next five to ten years, generative AI poses many potential ramifications for the legal market—including with respect to the training of young attorneys, the nature of legal work and how it is performed, and the structure and size of law firms. These potential ramifications should be carefully considered, particularly in connection with potential rule changes.

To begin with, generative AI will reduce the human capital needed to handle cases involving large quantities of documents, many parties, and complicated claims. Generative AI, thoughtfully applied, appears poised to improve upon existing TAR techniques[4] while also lowering expertise barriers for its effective

4. To be clear, improvement should not be taken as a matter of faith—this is an area where substantial independent study and evaluation are needed. Academics and practitioners should work

use. Likewise, generative AI, thoughtfully applied, seems likely to improve on the state of the art in legal research and improve efficiency. Drafting assistance is also likely to reduce significantly the number of hours that need to be spent performing certain drafting tasks. These enhancements create the possibility for certain legal services to be delivered more cheaply and broadly.

Second, these developments may alter the commercial landscape in the legal profession. How? We are not sure, but here are some plausible hypotheses:

- Some tech-savvy law firms, including small firms, will prosper, as they will be able to outbid firms with large overheads.
- Savvy large firms will figure out ways to capitalize on their expertise, size advantages, and intellectual property.
- Some internal corporate law departments using AI will be able to do more work, and more cheaply, than outside counsel, leading to an internalization of certain legal work that was previously sourced to outside counsel.
- Firms will need to compete over the best tech experts and technically sophisticated lawyers.
- Alternative fee arrangements are likely to grow more popular. The billable hour is likely to remain common, but firms will become increasingly eager to offer alternative billing structures to clients.
- More legal information and advice will become available online. There are statutory and ethical limits to the "practice of law," but under current rules there is leeway for information websites, for law firms to show their expertise with interactive websites for potential clients, and for clinics and legal aid firms to use AI to help potential clients.
- RoboLawyers? Yes, they are coming.

This is surely a partial list but ought to provoke further thought.

Generative AI also has substantial implications, which cut in many directions, for dignity across the legal profession—the dignity of the client and the client's relationship with their lawyer, the dignity of litigants lacking access to a lawyer's representation, and the dignity of the judiciary and court system. Accordingly, the ethical regimes governing lawyer conduct will grow more important for lawyers to know and contemplate.

Lawyers who use generative AI need to make sure that their clients are informed of, and consent to, the firm's use of technology, including AI techniques, in connection with the delivery of legal services and how data may be used by the firm to

together to determine whether and to what extent language models can match the existing state of the art in AI-assisted document review.

improve the quality of its services consistent with the ethical rules. Technological change thus makes it more important that counsel understand and follow the ABA and state bar rules concerning client communications and consent, confidentiality of privileged communications and client trade secrets, and client rights to the files and information developed in the course of representation.

As to the dignity of litigants and courts, AI has far-reaching implications. Generative AI may exacerbate imbalances in the adversarial system, overwhelm courts, and enable new forms of improper behavior by litigants. Courts may face greater pressure to police subtle litigation misconduct even as they find themselves managing ever-larger dockets. One also should keep in mind that procedural rules are tailored to certain assumptions about how much time, expense, and effort procedural maneuvers require. Generative AI will put pressure on these assumptions in unexpected ways.

At the same time, generative AI, responsibly used, has the potential to address major gaps in access to justice in U.S. society. A large majority of litigants are lawyerless. Many people and institutions have valid legal claims but cannot pursue them because they do not know their rights and/or lack access to lawyers who can press their rights. Generative AI has the potential to aid efforts to improve the status quo in this regard. Legal aid offices, innocence projects, law school clinics, legal services nonprofits, and for-profit law firms each have the potential to use new technology responsibly to expand the scope of their services, serve more clients, and charge less.

As with all things touching generative AI, however, such expansion should be undertaken carefully. The Model Rules' foundational requirement, zealous representation of a client, challenges all of us to think about how we can internalize our client's preferences. While technology represents a major opportunity to improve access to justice and meet the unmet demand for legal services, it also poses important questions about how lawyers can preserve the dignity of the attorney-client relationship as they represent many more clients.

III. Generative AI Will Change Legal Education and Scholarship

As members of the bar and as scholars, many legal academics will grapple with generative AI over the next decade. More specifically, however, generative AI has the potential to transform legal education and scholarship.

To begin with, the content of law school courses will change to account for new technological advancements. How can a civil procedure course not consider

the ways that AI can be used to limit the burden associated with document review? So too courses on property and copyright should confront cutting-edge issues raised by generative AI. Statutory interpretation is now a required or widely subscribed course at most law schools. Generative AI raises important questions for these courses, as well as potentially providing new tools to perform research into legislative history, original public meaning, and the use of statutory words and phrases throughout the law.

There is also the potential for law school pedagogy to change. The pandemic showed us how law professors can adapt to technology (e.g., Zoom classes), and generative AI will invite teachers to bring new tools into their classrooms. Law professors and law schools should embrace this opportunity to introduce students to new technology in a thoughtful way. As one example, in fall 2024, we expect to collaborate with a first-year procedure course that will incorporate generative AI into students' research, writing, and analysis.

Generative AI also offers opportunities for clinical education. Law school clinics often help people who would otherwise be lawyerless and give them redress in an often-alienating system. Properly used, generative AI has the potential to enable clinics to process more cases while empowering students to learn how to use technology more effectively and grapple with many teachable moments, and questions about the future of legal practice, under the instruction of clinical faculty.

Legal scholarship has already responded with an increasing number of law review articles, policy papers, and law school centers to study and report on the new technology. From the perspective of scholarship, generative AI seems poised to require scholars to reconsider central assumptions about many areas of law. At the same time, generative AI seems likely to provide academics new tools that can help them better study and understand the legal system.

IV. Productive Roles for the Legal Profession

There is substantial capacity for the members of the legal profession—including academics, bar associations, courts, firms, lawyers, law students, legal service organizations, and technology specialists—to shape the trajectory of generative AI in the profession.

In our view, the legal profession should engage in at least three kinds of activities to help its stakeholders become acclimated to generative AI.

First and foremost, the profession needs to prioritize evaluating the capabilities of generative AI models. The ABA and state bar associations can help smooth the way by providing ground rules for academic–practitioner collaborations that

advance the evaluation of generative AI in legal contexts. There is also significant untapped potential at law schools and firms to study different generative AI tools and publish (consistent with ethical responsibilities) the results of those studies. These evaluative efforts will clarify matters considerably.

Second, the profession needs to prioritize the education of lawyers. While there is substantial hype, misinformation, and well-meaning inaccuracies regarding how LLMs work, there is also a growing body of accurate, understandable material that can be provided for educational purposes. Professional bodies and law school faculties should promote this material and reject technically inaccurate material so that every lawyer can have an accurate foundational understanding of how these tools work. In addition to promoting educational material, law schools, law firms, and the ABA and state bar associations should also offer hands-on training sessions, where lawyers and law students can learn how to use and interact with existing technologies.

Third, the ABA, and state bar associations, should seek to clarify that existing rules of professional conduct already govern the use of generative AI and that generative AI may be used consistent with existing ethical requirements. As an example, bar associations could clarify that the duty of technological competence extends to the responsible use of generative AI–based systems—you can use the technology, but you need to understand it and implement appropriate safeguards to ensure its use is consistent with existing ethical and procedural standards.

V. Conclusion

These are interesting and exciting times for the legal profession. The technologies underlying the recent wave of interest in AI—large generative language models—remain flawed and unreliable in significant ways, with substantial unrealized potential. LLMs and the software systems built around them will improve in the years ahead, probably quite rapidly. Even if LLMs hit a wall, their usefulness will grow significantly as the profession understands them better and develops creative, novel ways to weave language models into existing workflows. We should expect significant change over the next five to ten years, particularly as to the day-to-day realities of practicing law. At the same time, the profession shouldn't lose its head.