7 Questions Lawyers Should Ask Vendors About Their AI Products

By Maura R. Grossman and Rees W. Morrison

The frenetic and much-touted world of artificial intelligence (AI) has poured into the legal industry like a storm surge. Lawyers who lack technical expertise or feel overwhelmed by jargon and arcane mathematical concepts are at a distinct disadvantage in this technology-oriented new world. Vendors can make assertions with little risk of cross-examination.

If your law firm or department has invited a vendor to explain or demonstrate its AI software, you likely already know the foundational questions to ask about the vendor's company, competitive position, pricing, support, and user base. These days, you likely also know to ask about the vendor's data protection and data security practices. However, you are probably on less solid ground concerning the questions to ask about the underlying machine-learning software. This article proposes seven basic questions – and a framework for understanding the answers to those questions – that are specifically targeted at vendors that offer AI and machine-learning products and services.[1]

Maura R. Grossman, J.D., Ph.D., is a Research Professor and Director of Women in Computer Science in the David R. Cheriton School of Computer Science at the University of Waterloo, and an Adjunct Professor at Osgoode Hall Law School of York University, where she teaches interdisciplinary courses on *Artificial Intelligence: Law, Ethics, and Policy*. She also is Principal at Maura Grossman Law, an eDiscovery law and consulting firm in New York. Maura's scholarly work on technology-assisted review (TAR) has been widely cited in the case law, both in the U.S. and abroad. She has served as a court-appointed special master, mediator, and eDiscovery expert to the court in multiple high-profile U.S. federal cases. Websites: http://grossman.uwaterloo.ca and https://en.wikipedia.org/wiki/Maura_R._Grossman.

Rees W. Morrison, Esq., a Partner at Altman Weil, Inc., has for three decades consulted to law departments and law firms on management issues, data analytics and recently on machine learning. Rees has published 320 articles and a recent book on visualizing and analyzing law firm data. He has written extensively about 500+ surveys sponsored by law firms on his blog JurisDatoris.com. A member of the College of Law Practice Management, Morrison runs the LinkedIn group, Law Department Management, with more than 3,200 members.

1. WHAT DO YOU MEAN WHEN YOU SAY YOUR SOFTWARE USES "ARTIFICIAL INTELLIGENCE" OR "MACHINE LEARNING?"

A subcategory of artificial intelligence, machine-learning software finds patterns in data, and the software improves its performance (i.e., "learns") as it processes more data. Data can include the words in documents – such as those contained in emails in electronic discovery or in word-processing files in contract analytics – which are analyzed using natural language processing or statistical methods. Data can also include figures from time and billing systems, where regression and neural networks can provide insights. Or data may be derived from human resources files, where classification methods, such as support vector machines or decision trees, can help identify records of interest or improve the quality of predictions.

The vendor should explain whether their software uses supervised or unsupervised learning. If supervised, your data will need labels (corresponding to classes or categories of interest, such as whether the client is a public or private company, whether the documents are privileged or not, or whether the practice group of a lawyer is corporate, litigation, or tax). In unsupervised learning, such as k-nearest neighbor classification, the software detects patterns on its own, based on the numbers in the variables.

What you should not hear from the vendor are grand, vague assertions, or that they cannot answer your questions because their software is based on proprietary methodologies.

2. HOW MUCH WILL WE HAVE TO CLEAN OUR DATA FOR IT TO BE USED BY YOUR SOFTWARE?

Almost always, machine-learning programs require the data that they process be in an organized format, much like a spreadsheet (for example, if the data consists of

numbers, it is called a matrix). Typically, the data will be stored and presented in rows and columns.

Before the software can run reliably, your data will need to be cleaned, for example, by making sure that the columns of data do not mix numbers and text or that they do not have missing values. Sometimes the software can handle missing values, but other times you may need to impute a value which reasonably estimates the missing value. You will also need to make sure that the codes you use for labeled information are consistent; for example, the names of courts need to be in a standard format.

Another reason the software needs properly prepared data is that the most common machine-learning methods depend on linear algebra – powerful mathematics that multiplies and manipulates matrices – and possibly also calculus and trigonometry to draw inferences from the data and optimize the output. If your data are untidy, the program will typically falter or fail.

Under ideal circumstances, you should not have to pay the vendor extra to pre-process your data, and you have (or can assemble) the necessary data in the requisite format to be read by the software. If that is not the case, you will need to figure out how the pre-processing will be accomplished and include that time and cost in your budget.

3. WHAT AMOUNT OF DATA AND TRAINING DO WE NEED TO USE YOUR SOFTWARE EFFECTIVELY?

The vendor should realistically estimate how many observations you need (think rows in your spreadsheet, or numbers of documents) and how many pieces of information you need about each observation (referred to as variables). With machine learning, more data is almost always better, but law firms or departments hardly need to have Big-Data volumes to be able to derive useful insights using machine-learning tools.

Regression, neural nets, and other machine-learning tools create a model from the data you supply. Typically, you provide the software with a portion of your data, the training set, and then vet the results of the model on a validation set, before you finally try the model on a hold-out or testing set to determine how accurate the model is. Your goal is to avoid overfitting the model so that it hews closely to the training data, but cannot take on new data and do a good job of classification or prediction.

What is important to understand is not only how much data will be needed, but also how much training on the software itself will be necessary before the software works properly. Most vendors will not reveal, without pressing, that it is uncommon for their software to work immediately, off-the-shelf, on your data, without additional training. You need to know how much tweaking or customization will be necessary so that you can add that time and cost into your assessment.

4. WHAT ALGORITHMS AND ASSUMPTIONS DOES YOUR SOFTWARE RELY ON?

You should push the vendor to explain clearly the algorithms and assumptions that underlie their software. Algorithms include if-then rules or instructions (e.g., "minimize this value") in the software code of the vendor's program that convert data into output or answers. In essence, they are the recipes that accomplish the classifications, conclusions, or predictions. Furthermore, it is important to understand the features the algorithm is using – such as age, gender, race, and so forth – so you are aware of underlying biases that may be hidden from your view.

You should also understand the concept of "hyperparameters." As previously mentioned with respect to training the software, hyperparameters are akin to knobs for tuning the machine-learning software, such as higher-level decisions about the learning rate of the process, or how significantly the software will adjust calculations called weights (in neural networks) or the loss function (in regression, where the most common choice is called "ordinary least squares"). The bottom line is that more knobs mean more nuanced learning, but also more complexity. In the same way that an automatic transmission is preferable to stick shift for most drivers, so too, extensive knob twiddling may require data science expertise the firm or department will need to obtain.

What you want to avoid is proprietary algorithms that are black-box and hard-coded so that your understanding of their inner workings is limited and your flexibility to match the software to your data and needs is constrained.

5. WHAT RESOURCES WILL WE NEED TO IMPLEMENT YOUR SOFTWARE SUCCESSFULLY?

Many implementations of standard machine-learning algorithms are available. Free open-source software packages like LibLinear and Vowpal Wabbit apply these algorithms to a spreadsheet-like representation of the processed data. Many popular programming languages provide access to implementations of these algorithms through the use of a computer program. Among the most popular languages, Python and R are free and open-source, while others, like SPSS, SAS, Stata, and MatLab, are proprietary. Many vendors in the legal space offer machine-learning tools, some of which use the machine-learning implementations described above and some of which are vendor-specific. You need to know how widely

available the people and resources are that can use the vendor's particular software.

You generally do not need to know that much about hardware, since the relatively modest sizes of most legal data sets should not require specialized capabilities or power such as graphical processing units (GPUs). However, you may need to drill down on vendors who do computations and storage on a cloud server, such as Microsoft Azure or Amazon Web Services, for example, if you are handling huge electronic discovery datasets. With cloud providers, issues concerning data protection and data security will need to take more prominence.

6. WHAT TOOLS DO WE NEED TO INTERPRET THE MACHINE-LEARNING MODEL AND TO VISUALIZE IT, AND ARE THEY INCLUDED WITH YOUR SOFTWARE?

Data scientists have created a range of tables, decision trees, and graphs that can help users probe and understand the insights to be drawn from their data. Tools can display in different visual formats the calculations performed by the machine-learning algorithm and the results they produce.

You should ask the vendor to explain and show you the

You will need to have someone available on your staff – or hire someone – to help navigate through data preparation, running the software, and, perhaps most important, interpreting the results. These individuals are typically referred to as data scientists. As just one example, machine-learning software often works better when the data has been normalized, i.e., all the figures, such as collections per office, are converted into a standard scale between 0 (for the least) and 1 (for the most); someone needs to understand whether and how to normalize the data and then how to interpret the output.

tools they make available for graphical analysis, interpretation, and display of results. Further, the vendor should show you what typical output will look like so that you can assess how interpretable the software's results are. If the vendor is using a neural net (or a stack of neural nets, which is referred to as "deep learning"), the vendor needs to explain how much of their software's effectiveness lurks in a black box. If you cannot figure out how the algorithm achieved its results, it may not be the right tool for you, especially if you have to

explain the output to your clients, or to your adversary or the court in litigation.

7. HOW HAS YOUR TOOL BEEN VALIDATED FOR ITS INTENDED PURPOSE AND HOW RELIABLE IS IT?

Finally, before you license the tool, it is imperative to know what empirical support there is that the software you are about to purchase is valid and reliable. Has independent testing or verification been performed? By whom and on what data? Asking for references from current users of the software is helpful, but less authoritative.

"Validity" refers to the extent to which the tool measures what it is supposed to measure; the extent to which the input is relevant to the output being assessed, and the extent to which responses on a measure can accurately classify or predict future behavior. "Reliability" refers to the extent to which the tool yields the same results over multiple efforts.

The vendor's tool should be provably valid and reliable. Just because a vendor claims that their tool is 99 percent accurate does not mean that it will work for your intended purposes, particularly if your situation is substantially different from the use on which the tool was tested. For example, it is easy for a vendor to claim that a tool is 99 percent accurate in predicting privilege, if only 1 percent of the data is privileged. The tool can misclassify 100 percent of the privileged data by labeling every document in the collection as non-privileged and still be 99 percent accurate. Do not be fooled by claims that do not consider both false positive and false negative errors. Make sure you understand what testing has been done to demonstrate that the software works and works consistently, and better yet, demand a proof of concept and do a test run yourself so you can *vet the tool on your own data to make sure it works as promised.*

CONCLUSION

While the questions above do not represent *all* of the questions a lawyer considering an AI product should ask a vendor, the answers to these seven questions will put you well on your way to (1) making sure that you have a good grasp of the product you are purchasing, (2) understanding the choices your firm will need to make when you use the vendor's software, (3) accounting for the additional help you may need (and will have to pay for) to use the tool effectively, and (4) avoiding unnecessary professional or reputational risk.

Of course, equally important aspects of the AI vetting process, beyond the scope of this article, include clearly identifying the problem that needs to be solved, making sure the proposed solution addresses that problem, and assessing that the proposed solution will work as expected in your unique environment.

1. Specialized jargon abounds in the field of machine learning. At minimum, you should probably familiarize yourself with terms such as "regression," "neural net," "support vector machines," and "deep learning," as well as basic statistical concepts. A useful glossary of technical terms primarily but not exclusively related to electronic discovery can be found at Maura R. Grossman and Gordon V. Cormack, *The Grossman – Cormack Glossary of Technology-Assisted Review*, 7 Fed. Cts. L. Rev. 1 (2013), http://www.fclr.org/fclr/articles/html/2010/grossman.pdf.