Trusting AI Systems

James W. McPhillips, Elizabeth Zimmer, Sandro Serra, and Mia Rendar*

In this article, the authors discuss artificial intelligence (AI) systems and risks in commercial contracting, the legal risks of AI systems in technology services, and the need for "explainability" in AI systems.

Artificial Intelligence Systems and Risks in Commercial Contracting

Although the use of artificial intelligence (AI) has grown steadily during the past decade, the recent release[1] of OpenAI's generative AI system, ChatGPT,[2] has led to a precipitous increase in attention and publicity accompanying the rise of powerful generative AI systems.

With these generative AI systems come mounting issues and concerns around the use of AI systems by technology service providers.

What Is an AI System?

To demonstrate the function of these generative AI systems, we asked ChatGPT to answer the question, "What is an AI system, and how does it work?" Its response:

> An AI system is a computer program that can perform tasks that normally require human intelligence, such as understanding natural language, recognizing objects in images, or making decisions. Generative AI is a type of artificial intelligence that involves the use of algorithms to generate new data, such as images, text, or audio, that has not been seen before.

If we were to provide this query to a general search engine, we might be provided responses that include the term "AI system," and would then review each resource and compile a response based on the information provided. The difference between conducting a general search, and using the generative AI, is that the generative

Robotics, Artificial Intelligence & Law / July–August 2023, Vol. 6, No. 4, pp. 247–255.

ISSN 2575-5633 (print) /ISSN 2575-5617 (online)

AI compiles the response for you, complete with correct syntax and wording responsive to the question.

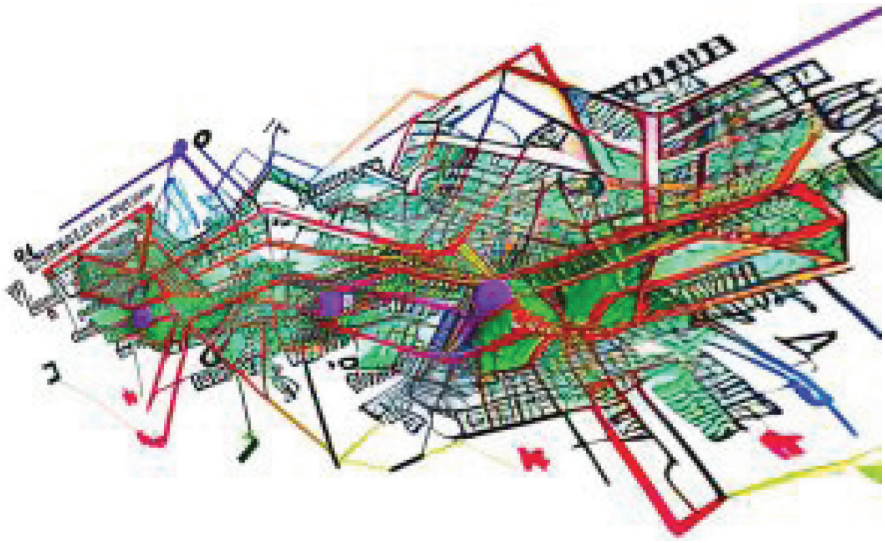
How Does an AI System Work?

AI systems work by analyzing large amounts of data and extracting patterns and insights from it. The analytical system can learn to recognize patterns in data and use those patterns to make predictions or classifications about new data. In other words, AI is basically an incredibly complex and exceedingly large decision tree, powered by statistical probability.

More technically, these decision trees take the form of Markov chains,[3] or stochastic methods. A Markov chain is a mathematical model that describes a system that transitions between different states over time. It is a stochastic process, which means that the next state of the system is determined by a probability distribution that depends only on the current state of the system and not on any prior states. In natural language processing, Markov models are often used to model the probability distribution of words in a text. For example, a first-order Markov model would predict the probability of a word based on the probability of the previous word. A second-order Markov model would predict the probability of a word based on the probability of the previous two words.

Markov chains (see Figure 1) provide a powerful framework for modeling and understanding sequential data in machine learning and AI applications. But it is important to note that AI systems need a significant amount of data to "train" the algorithm, or derive the probabilities necessary to create the chain. In the case of a language model like ChatGPT, the application has been trained on vast amounts of text data and has learned to generate natural sounding language by predicting the most likely next word or phrase based on the previous context. ChatGPT touts that it draws its material from a wide variety of sources and domains, including books and literature, web pages and articles, and social media and messaging. Similar to the repetition required to train a dog, an AI system must experience repeated patterns in order to "learn" to provide the required result. The source of this training data can be a hot topic, and worth considering as AI becomes more and more prevalent.

Figure 1. Markov Chains (Modified by Generative AI)



What Are Some Examples of an AI System?

We have grown accustomed to AI systems like "Siri" and "Alexa," both of which can be fed queries and requests, and in turn select responses and complete tasks from a closed list of possible responses. Other familiar AI uses are the text-completion function in certain email products, image recognition in popular social media sites that suggests location or personal tagging, and autonomous driving or self-driving mechanisms in cars. And, of course, we have grown accustomed to AI chatbots in a number of contexts.

Unlike AI bots of yore, ChatGPT's responses remember context of the ongoing conversation and can be prompted to perform deeper or second-level instructions, like providing a response that fits a certain style, or using certain defined terms. These generative capabilities have enabled a number of new players on the technology scene around AI, but also some more familiar service providers are developing competitive AI systems.[4]

As these AI systems become more common in business settings, the reality is that using this technology is not without risk. This article next examines the legal risks of AI systems in technology services.

The Legal Risks of AI Systems in Technology Services

The first part of this article provided an introduction to the budding new technology of generative AI, or AI systems. As with the implementation of any new technology, widespread understanding of the risks generally lags behind the speed of the technology itself. When the technology industry began its push "to the cloud," many customers were concerned about certain issues, including, but not limited to, giving up control of data, security risks, and performance issues. In response, sophisticated customers carefully addressed these types of issues in their contracts with cloud service providers.

A similar approach is likely to play out with respect to AI technology. The market will ultimately drive how AI risk is addressed, but at the moment, we see several risks and issues for AI adopters to consider carefully, discussed below.

Confidentiality and Security of a Customer's Data

A common term in commercial contracts generally, and technology service provider contracts specifically, is an obligation for each party to maintain the confidentiality of data or information provided as part of the engagement. In addition, and particularly for customers in heavily regulated industries, the most robust confidentiality terms are imposed on service providers that have access to or host a customer's data.

Interfacing with a service provider using AI should be no different. With respect to contractual protections, customers should ensure that AI service providers agree to meet appropriate obligations (i.e., both traditional confidentiality terms as well as more robust technical security requirements) to protect the confidential nature of data and information. Customers should also look out for how "customer data" is defined and ensure that all data it provides the service provider is subject to the confidentiality and security obligations, including information derived from the data it provides as part of the engagement.

Risk can also be mitigated outside the contract. For example, customers should consider implementing internal procedures that limit exposure, such as restricting users from sharing personal or

proprietary information, or requiring encryption or other means of security prior to the data ever reaching the AI system.

We recommend that customers review their current engagements with AI providers to ensure that (1) "customer data" includes all of the data, information, and materials a customer provides to the service provider, as well as all materials derived from such data, and (2) the confidentiality and security obligations clearly apply to all such data processed via an AI system.

Commercial Value of Customer Data

In addition to protecting confidentiality and security of data, customers should be careful about protecting the commercial, proprietary value of its data and the derivatives of such data. AI products use huge amounts of data to learn and improve their models. If a customer owns the input data, and such data has commercial value, then a customer may want to restrict how service providers use such data to improve their AI products. That said, the improved models provide much of the value of an AI product, so service providers will also likely negotiate this issue heavily, and such negotiations can be rather complex.

Customers purchasing AI products should consider including express contractual terms where the customer retains ownership of all preexisting materials. In addition, customers should establish a clear position as to how service providers are able to use customer data.

Third-Party Liability

As noted above, AI systems learn from a wide variety of data sources. Service providers selling and licensing these systems must have the appropriate rights and consents to use data from all of these sources. If an AI service provider has not secured the appropriate rights to use the information, an individual customer could be exposed to risks of infringement or misappropriation from a third party for the customer's downstream use.

There is still a great deal of uncertainty around how the generative AI tools available for public use are handling scraping of proprietary information. On November 3, 2022, a class action lawsuit[5] was filed against a number of AI system service providers,

asserting that the providers' scraping licensed code to create AI-powered tools was a violation of licensing terms applicable to code repositories. The suit was dismissed for lack of injury and failure to state a viable claim, but the rumblings related to how these companies are leveraging data they scrape from "public" sources is worth noting.

Because of the lack of certainty, customers should insist that an AI provider bears the risk associated with the customer's use of the AI system. A customer should consider including indemnity obligations that cover third-party claims associated with intellectual property or privacy violations, and ensure liability for such claims is not limited unduly by a cap on damages.

Regardless of the above risks, contracting for technology services that include AI systems implies that we can trust AI systems to effectively perform the tasks we want them to. Next, this article explores the risks around the efficacy and accuracy of AI systems.

Earning Your Trust: The Need for "Explainability" in AI Systems

AI systems seem like an exciting, effective new tool. But, as we have seen with Google's recent struggles with accuracy,[6] and Microsoft's trouble with sentient, unhinged chatbots,[7] not all of the kinks have been worked out with these tools.

The previous part of this article discussed the legal risks, and related contractual mitigants for entering into agreements with AI vendors, but perhaps a more pressing question is whether one can trust AI systems in the first place.

Bias and Reliability

As some say, you are what you eat. AI systems eat up immeasurable amounts of data, and ultimately, AI output and results are only as good as the inputs they process. If the inputs are unreliable, or biased in any manner, they will invariably result in biased or unreliable outputs.

As mentioned above, we know that the decision making is based on training data, overlaid with probability-based decision making. Resultantly, errors or idiosyncrasies inherent in the training data lead to cumulative errors in the results.

For example, an AI system used to screen job applications may inadvertently favor male applicants over female applicants if the system was trained on a data set that contained more résumés from men than from women. Similarly, a facial recognition system may have higher error rates for people with darker skin tones, as the training data may have included fewer examples of darker skinned individuals.

Another example is an AI system queried to provide insights on the risks of certain medical procedures,[8] and pulling resources from chat pages or message boards that provide unqualified or misinformed advice on such procedures.

AI bias and unreliability can have serious consequences, particularly in applications such as medicine, hiring, lending, and criminal justice, where biased or faulty decision making can perpetuate discrimination or misinformation—and could result in running afoul of fraud or discrimination laws.

Mechanisms to avoid receiving AI services that are unreliable or discriminatory can be legal or operational. Some strategies to operationally mitigate the risk include having• open conversations with the AI provider to discuss their efforts to avoid AI bias and unreliability. Generally, understanding how the AI technology functions (including if inherent bias exists or if the data inputs are vetted), or if the AI provider has not fully accounted for procedures to avoid bias or misinformation, allows customers to implement their own compliance mechanisms to close the gaps themselves, and make more informed decisions to reduce or avoid the possibility of bias or unreliability.

Some elements a customer might consider building into its contract with an AI provider are:

1. Clear descriptions of the AI system specifications, including non-discriminatory and fact-checking features and practices;
2. Representations and warranties that shift the burden of proving that discrimination or fraud did not occur to the AI provider; and
3. Indemnification obligations requiring the AI provider to cover claims that the AI system caused discrimination or were factually incorrect.

We recommend operationalizing internal controls as well as developing legal standards that address the above points.

Transparency and Explainability

Many AI systems are considered "black boxes," meaning their decision-making processes are not transparent. This can make it difficult to understand why the AI system is making certain decisions or predictions, and it can be challenging to identify errors or detect problems.

"Explainability" is becoming increasingly important as AI systems are deployed in high stakes applications. For example, if a self-driving car causes an accident, having the ability to determine why the car made the decision it did and whether the decision was reasonable prevents future accidents or errors.

In addition to helping with accountability and transparency, explainability can also help developers and researchers to identify errors, biases, and other problems with AI systems, and to improve the accuracy and reliability of the models.

To improve explainability, researchers and developers are exploring various techniques, such as creating models that are more transparent and interpretable, developing algorithms that can explain their decisions in natural language, and using visualization tools to help users understand how the AI is working. Feature importance analysis[9] identifies which input features are the most important for the model's predictions, and decision rule extraction,[10] as the phrase suggests, extracts decision rules from the model.

Admittedly, measuring the explainability of an AI system can be subjective, as it often requires human interpretation. One approach is to use surveys or user studies to evaluate the interpretability of the model. Another approach is to use complexity metrics such as the number of parameters or the size of the model to measure the complexity of the model.

Customers utilizing tools may have difficulty controlling for AI explainability. To mitigate the risk, customers should consider requesting references from its AI service providers to learn from other customers whether the AI services are functioning in a clear and transparent manner, implementing frequent testing of results to ensure a human is assessing the quality of the output. More broadly, given the burgeoning uses of AI systems, it may not even be transparent when AI systems are actually being used. To avoid leveraging products that are unknowingly subject to the risks and issues we have identified related to AI systems, consider requesting

clearer technical descriptions of products and any machine learning that might occur. Also consider preemptively building standard machine learning and AI requirements into your technology and professional services master agreements.

While AI systems hold immense promise, their risks and limitations cannot be ignored. Bias, unreliability, and lack of transparency are just some of the issues that need to be addressed when considering the use of AI systems. It is important for customers to have open discussions with AI providers about their efforts to mitigate these risks and to understand how the technology functions. By taking these steps, customers can reduce the possibility of bias or unreliability, promote accountability, and transparency, and ultimately make more informed decisions about the use of AI systems.

Notes

* The authors, attorneys with Pillsbury Winthrop Shaw Pittman LLP, may be contacted at james.mcphillips@pillsburylaw.com, elizabeth.zimmer@pillsburylaw.com, sandro.serra@pillsburylaw.com, and mia.rendar@pillsburylaw.com, respectively.

1. https://www.wsj.com/articles/chatgpt-ai-chatbot-app-explained-11675865177.
2. https://chat.openai.com.
3. https://en.wikipedia.org/wiki/Markov_chain.
4. https://www.nytimes.com/2023/02/07/business/dealbook/chatgpt-google-baidu.html.
5. https://www.reuters.com/legal/litigation/openai-microsoft-want-court-toss-lawsuit-accusing-them-abusing-open-source-code-2023-01-27/.
6. https://www.npr.org/2023/02/09/1155650909/google-chatbot--error-bard-shares.
7. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html.
8. https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html.
9. https://towardsdatascience.com/understanding-feature-importance-and-how-to-implement-it-in-python-ff0287b20285?gi=60ce54363555.
10. https://ieeexplore.ieee.org/document/938448.