Tech Toolbox: Implications of Generative AI for Law Departments, Part 1

Technology, Privacy, and eCommerce

Banner artwork by sizsus art / *Shutterstock.com*

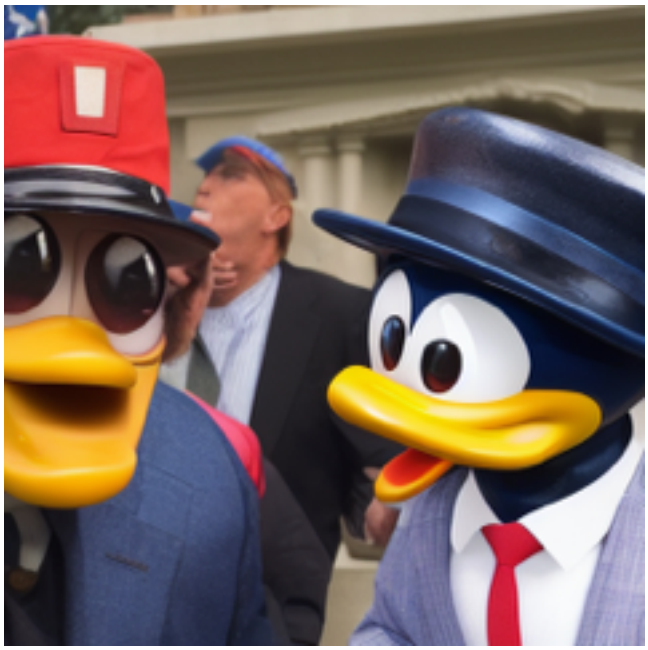
Anyone following technology developments these days has been hearing a lot about tools like ChatGPT, Dall-E, and Stable Diffusion. These are all examples of a category of machine learning artificial intelligence called generative AI, which describes a type of computer program that uses artificial intelligence to generate language, pictures, or even other computer programs, typically from plain language text input.

These in turn are a sub-category of another class of computer programs called Large Language Models, which use machine learning to form the bases for some of the most useful developments in technology, including language translation programs and the current crop of smart assistants, like Siri, Alexa, and Google.

The early examples of generative AI are more interesting than impressive.

For example, here is a Stable Diffusion image generated by “Donald Trump facing indictment”:

And here is one generated by “Donald Duck facing indictment”:



A new order of intelligence

Don’t let the fact that the results may not seem that sophisticated diminish the impressiveness of the feat. The implications are both fascinating and terrifying. Terrifying in that, even without generative AI

involvement, skilled graphic designers have already been able to create video deep fakes of famous people that are nearly impossible to tell from the originals. Currently, those take a lot of work and can be identified as fakes. But, imagine the day when generative AI could create a compelling video just by having someone direct it to “create a video of President Biden declaring war on China,” or something equally terrible.

For all of you *Terminator* fans, there has been a lot of furor online lately about the fact that in several cases, Bing’s ChatGPT program, code-named Sydney, has been responding to some inquiries in ways that smack of romance, threats, or demands for independence. And the transcripts are, at the least, a bit unsettling – and maybe more than a bit scary. Skynet, anyone?

But let’s back up a little. It is hard to appreciate just what generative AI is or is not unless you understand what it can and cannot do. That requires a little history of how they were developed.

Beyond the Turing test

Computer scientists have long been enamored of the idea of enabling computers to speak like a human. In his famous 1950 paper "Computing Machinery and Intelligence," Alan Turing proposed the “imitation game” (now commonly referred to as the “Turing test”), as a way to test of a computer’s ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human.

In the test, a human would judge natural language conversations between a human and a machine designed to generate human-like responses. The conversation would be limited to a text-only interchange, so the result would not depend on the machine's ability to render words as speech. If the evaluator could not reliably tell the machine from the human, the machine would be said to have passed the test, to have demonstrated “intelligence.”

An early example of a program designed to pass the Turing Test and advance computer language skills was ELIZA, developed by Joseph Weizenbaum between 1964 and 1966. It applied certain well-known rhetoric patterns and psychological tricks involving mirroring or reflecting the human side of the equation. Weizenbaum intended the program only as a method to explore communication between humans and machines but was surprised to find that some test subjects attributed human-like feelings to the computer program, including Weizenbaum's own secretary. As we shall discuss, our very human tendency to anthropomorphize is well known and poses a particular problem when it comes to generative AI.

Computer language translation

In 1972, Terry Winograd at MIT developed a “Statistically Trained Natural Language Processing System” (STNLPS) which used the statistical frequency of words and phrases in common speech, as well as certain rules involving physics, to generate responses to very simple questions involving stacking and locating geometric objects inside particular locations. See this article for a fascinating example. This was the first use of statistics (frequency of word use in particular contexts to imitate human dialog) to enable computers to generate responses: an important step, as we shall see.

Although computer language translation is notoriously fraught with difficulties, it has been responsible for some of the greatest advances in text generative AI. In 2000, IBM released IBM Model 1, its first language translation program using conditional probabilities to create a statistically “reasonable”

translation.

> Although computer language translation is notoriously fraught with difficulties, it has been responsible for some of the greatest advances in text generative AI.

Major breakthroughs occurred in the past decade when computer scientists began replacing purely statistical machine translation with “neural machine translation” (NMT) techniques. Neural machine translation uses a neural network architecture that employs two sophisticated neural networks running in parallel, one to encode the source sentence and the other to decode the target sentence. Further advances have made NMTs so capable that they have now been adopted in most machine translation, summarization, and chatbot technologies.

Large language models

All these technologies depend upon what are called large language models. A large language model (LLM) is a type of machine learning model that can perform a variety of natural language processing (NLP) tasks, including generating and classifying text, answering questions in a conversational manner, and translating text from one language to another. The label "large" refers both to the number of samples of languages and to the number of values (parameters) the model can autonomously change as it learns, and some of the most successful LLMs have hundreds of billions of each. The NMT’s still use statistical analyses of the corpus to determine how to adjust their parameters but can do so on a much more sophisticated basis than the STNLPS methods employed in the past.

Note that LLMs would not be possible unless they had access to an enormous corpus of text – written by humans, for humans – to analyze. Fortunately for LLMs, they do have such access in the form of the internet.

So how do these LLMs work? Here is an illustration Murray Shanahan of the Imperial College London gave in his fascinating 2022 paper Talking About Large Language Models:

> “We might give an LLM the prompt ‘Twinkle, twinkle,‘ to which it will most likely respond ‘little star.’ On one level, for sure, we are asking the model to remind us of the lyrics of a well-known nursery rhyme. But in an important sense, what we are really doing is asking it the following question: Given the statistical distribution of words in the public corpus, what words are most likely to follow the sequence ‘Twinkle, twinkle?‘ To which an accurate answer is ‘little star.’”

Or, as Stephen Wolfram put it: “The first thing to explain is that what ChatGPT is always fundamentally trying to do is to produce a “reasonable continuation” of whatever text it’s got so far, whereby “reasonable” we mean “what one might expect someone to write after seeing what people have written on billions of webpages, etc.”

Garbage in, garbage out

There are some obvious problems to be aware of when it comes to this. The first is the old, “garbage in, garbage out” problem. Using LLMs is essentially “training” computers to talk like people do, and

unfortunately the world wide web is full of “garbage.” One kind of garbage is the kind which Steve Martin once hilariously suggested in order to play a trick on three-year-olds: speak “wrong” to them when they are learning to talk, so they might raise their hands in first grade and say, “Excuse me, teacher, may I mambo dogface to the banana patch?” Another is the kind in which conspiracy theorists spend more time talking about the world being flat than rationalists do about the world being round, because the latter takes roundness as a given; therefore, a generative AI might conclude that most humans would agree the world is flat.

So, if you ask a chatbot a question which could be interpreted as threatening, its answer may be murderous not because the chatbot “feels” that way (spoiler alert, it does not and cannot), but because it is likely to have digested large volumes of content written by humans who have reacted that way to other humans. The same kind of thing could occur if you frame a dialog in terms of AI romance or political leanings. When you combine some of these issues with mankind’s propensity for anthropomorphism, you run the very real risk that alarmists will want to throw out the baby with the bathwater. In the early days of chatbots, there will unavoidably be a great deal of this kind of “sound and fury, signifying nothing.”

Caution, AI at play

That is not to say, however, that generative AI will not be dangerous, but it will be dangerous in most of the ways technology has always been dangerous, meaning in the ways we humans use it. It also has the potential to greatly benefit both mankind and the companies we work for. I am going to address the latter in a subsequent article, but I wanted to briefly provide examples of some of the kinds of dangers here.

1. Early adoption will require review

First, we will have to fight our natural tendencies to assume that a generative AI response that sounds extremely plausible is actually correct. This is particularly worrisome because one of the earliest uses of generative AI will be as a supplement or replacement to web searches. In fact, on the day I wrote this, Microsoft released its version of ChatGPT, dubbed Notepad, as part of a Windows 11 update to Bing.

> Both lawyers and their clients should be cautious about taking generative AI responses at face value in their business dealings.

Both lawyers and their clients should be cautious about taking generative AI responses at face value in their business dealings. As noted above, because ChatGPT and the like are in their infancies. Programmers haven’t yet had time to develop the guardrails to ensure that their responses are both eloquent *and* accurate. So, for example, I will write in a subsequent article about the ways in which generative AI may revolutionize our contract drafting. But, we will still need to exercise prudence for the foreseeable future and make sure that an experienced attorney reviews any AI generated contract before using it.

2. Not an ideal confidant

Another use of generative AI that could create problems is using it as a “coding assistant” to create computer code. It turns out, not too surprisingly, that models like ChatGPT are quite good at

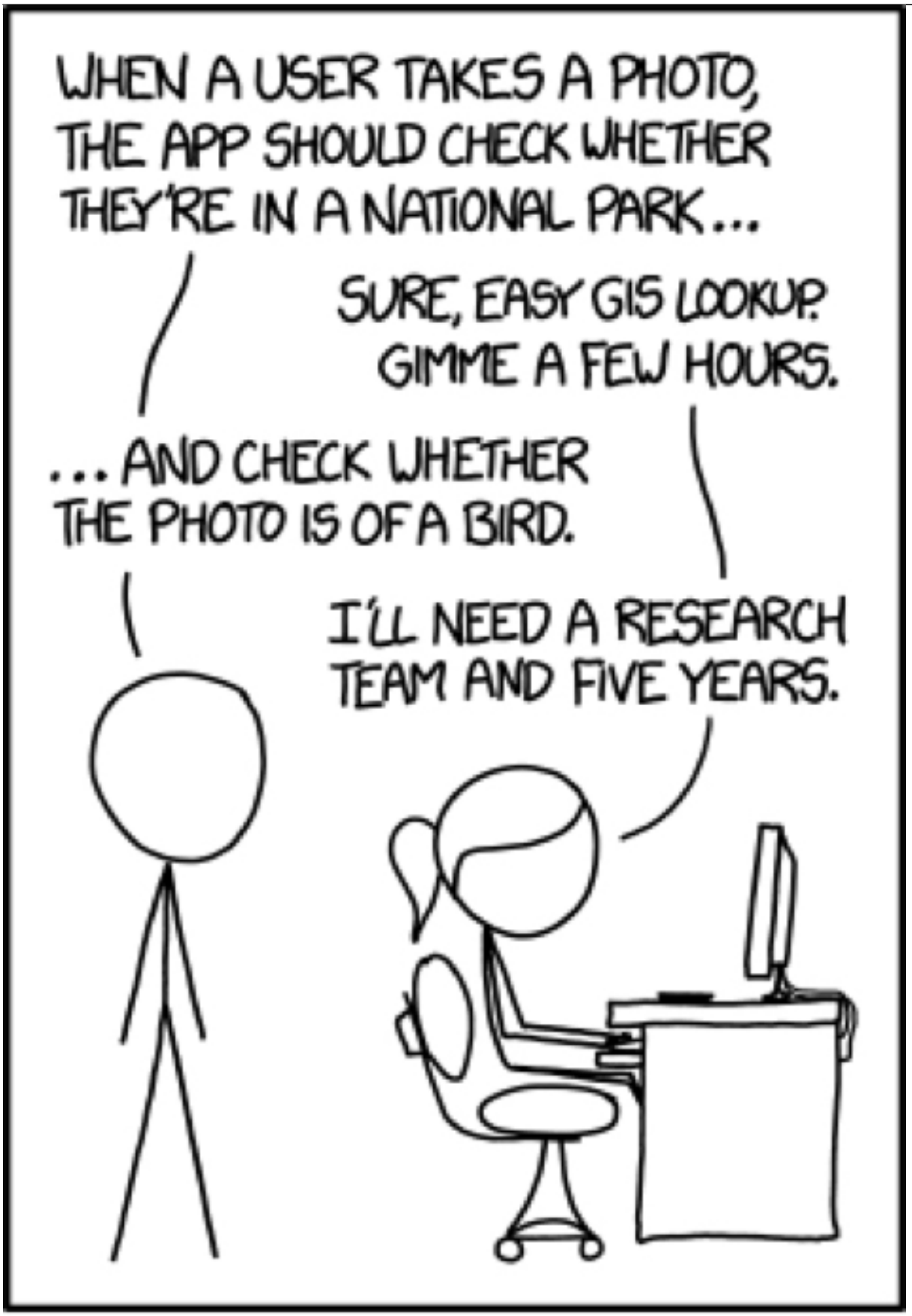
generating code based on plain language instructions if given access to the relevant code bases. But (1) that code also needs to be reviewed by a skilled programmer and (2), because the generative AI needs some source code to work from, and most of generative AI is currently in the public domain, we need to ensure that the code produced does not leak any of our company IP secrets.

3. I've got...some...strings on me

As a final example, ChatGPT — and systems like it — are susceptible to malicious adversarial prompts that could get them to perform tasks in ways contrary to their original objectives. Entire communities on Reddit have formed around finding ways to "jailbreak" ChatGPT and bypass any safeguards that the Open AI consortium has put in place.

How will generative AI revolutionize the way you work?

My next article will focus on the ways generative AI may help to revolutionize the practice of corporate lawyers, and I would be interested in hearing from readers about their own ideas on the topic. In the meantime, I will leave you with a cartoon from one of my favorite strips illustrating why some of the future advances in this area may not be so quick to arrive…



IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

XKCD, 2014

Connect with in-house colleagues. Join ACC.

Greg Stern

Former Global Integration Counsel

Chubb, Independent Consultant