Artificial intelligence, chatbots and the Bar

Ben Kremer SC
Banco Chambers

Artificial intelligence ('AI') is currently a hot topic. There are predictions that it will drive the next 'dot com'-like boom. Microsoft, Google and Meta (Facebook) are racing to release updated business- and consumer-facing AI platforms, while companies are moving equally as fast to integrate ChatGPT technology into their products. The longer-term outlook is less clear, and some see it as less than rosy; Henry Kissinger recently said in an interview that he was now trying to do with AI 'what I did with respect to nuclear weapons, to call attention to the importance of the impact of this evolution'.[1]

It is already clear that AI will have a big impact on the practice of law, including at the Bar, and that we are in the early stages of its adoption. However, what AI is, and how it works, is not generally well understood. Identifying three key aspects of AI may assist members of the Bar, as understanding what it is greatly assists in identifying what its benefits and pitfalls are likely to be for us in practice.

AI means many things

The first aspect is that AI is an umbrella term: it does not – when viewed at the technical level – describe a single, uniform thing, but rather many different types of technology. Computers can be programmed to solve narrowly defined problems, like winning games such as chess or Go; 'pattern recognition' problems, like understanding human speech or recognising images; or 'generative' problems, like creating new output of a particular kind or style. Each of these involves different programming techniques, and systems that excel at one task are often not easily generalisable outside it. Further, quite different techniques may be used to solve the same task.

For example, a chess AI is optimised to identify huge numbers of permutations of a finite number of known pieces that follow clearly defined rules on an 8x8 grid playing board. It may do this by generating 'trees' of permutations of future legal moves, which are then 'pruned' to trim ones likely to lead to a loss and leave ones most likely to win, adding new branches to the tree after every move.[2] Or it may do so by a 'genetic algorithm' that spawns multiple programs with subtly different evaluation functions; runs them all; discards all but the most successful 20 per cent; spawns replacement programs drawn from slight variations on those 20 per cent; and then repeats the cycle hundreds or thousands of times to leave the strongest, thereby mimicking Darwinian evolution.[3]

Conversely, a pattern recognition AI such as in a self-driving car may reconstruct a 3D data model of the real world by combining input from multiple cameras using a 'neural network', which uses software to replicate simple, interconnected nodes that work similarly to (but at a massively smaller scale than) the neurons in a human brain. Generally speaking, each 'node' takes input – whether from a camera or another node – applies an algorithm, and then produces an output.[4] The nodes are interconnected, and signals may pass through many nodes, being transformed each time. The combined set of outputs can, depending on the algorithms and the number of nodes, be interpreted to identify objects and give parameters such as size, distance, depth, or velocity.

These are only three of the common kinds of AI programming techniques, but it should already be obvious that one must have regard to what is happening 'under the hood' in order to assess the impact of any particular form of AI in a given context.

AI chatbots: not a new concept

The second aspect is that the current surge of interest in AI is heavily driven by 'chatbots'.[5] A user can type a natural language query into a text box and receive a natural language response that appears to come from the computer. The user can then respond, and the impression the user gets from interacting in the form of dialogue is that there is a thinking, knowing entity on the other side of the text box and that the entity is truly intelligent.[6]

This form of AI is actually not new; a chatbot program called ELIZA was developed in the mid 1960s, and versions for personal computers such as the Apple II appeared in the early 1980s. ELIZA was designed to produce output resembling the speech of a psychotherapist, and in short exposures it could produce very realistic-sounding conversations. However, it relatively quickly became obvious that its output was based on the user's input, and it was easy to induce repetition. The current phase of the AI boom has been sparked by modern chatbots such as ChatGPT, which can hold much more realistic-sounding conversations in which it can synthesise apparently new and creative content that is not directly derived from the user's input.

The reasons for the difference can be seen by applying the first point noted above: how did each chatbot work? ELIZA worked by disassembling the sentences typed into it and producing answers from pre-generated phrases triggered by, and often containing, different keywords in that input, along with stock phrases such as 'I see' or 'In what way?'[7] These were often in the form of questions, prompting the user to enter new content, which would produce different responses. The programming was brilliant for its time, given that it had to run on computers with less processing power than a modern microwave oven, but the constraints meant that the content it produced fell within a small, and eventually predictable, range.

By contrast, modern chatbots such as ChatGPT work very differently. They consist of a neural network that must first be 'trained', which essentially involves loading data into it in the form that the chatbot is to produce. So, a chatbot focussed on outputting computer source code is trained by feeding it source code, while a chatbot focussed on literature is trained by feeding it works of literature. General purpose chatbots like ChatGPT are trained on multiple inputs (e.g. books, Wikipedia articles, and web pages). The key point is that the source material used for training contains unique words, as well as language patterns in which those words are found.

In operation, the neural network essentially applies statistics: it analyses the query made to it, which sets up initial parameters, and then produces its output word by word. It chooses each successive word by looking at what has already been output, and then choosing a word according to the characteristics of similar patterns found in the dataset it was trained on. The choice between different possible words is, of course, highly complex, incorporating (for example) algorithms to try to avoid repetition or dead ends and to increase the probability of drawing upon source material related to the content of the user's query. But the method could be summarised by saying that the output 'once upon a' has a high chance of being followed by 'time' if the dataset contains fairy tales, 'midnight dreary' if the dataset contains poetry, or 'time in Hollywood' if trained on movies. Patterns and words found in the dataset will be replicated in the output, with words and patterns found more commonly in the dataset being output more frequently. Hence the reason this kind of AI chatbot is described as a 'generative pre-trained transformer' ('GPT'): broadly speaking, it generates output by using a 'transformer' (a term of art in machine learning) that differentially weights (using probabilities) parts of the pre-trained text to pick what word should come next.

Once this is appreciated, a number of points can be made. First, and perhaps surprisingly, this statistical method of producing output can result in realistic-sounding conversation and prose, as well as output that looks to exhibit creativity. Simply put, human language contains patterns that lend themselves well to a statistics-based generative approach.

Secondly, the output of a chatbot depends heavily upon the material it can draw upon: the larger the training dataset, the more different patterns and words can be drawn upon, and the more realistic and relevant the output appears. The rapid improvement in chatbots is in large part due to this: ChatGPT-1 was trained on 5 GB of material, ChatGPT-2 on 40 GB, and ChatGPT-4 on 570 GB. However, it can also be seen that the type of content produced will be limited by the dataset: a chatbot trained on source code will not be able to output prose, and a chatbot trained on Shakespeare will not be able to output *Hamilton.*

Source: reddit.com/r/ProgrammerHumor, creator unknown

The third and most important observation, however, is that although a human might perceive the chatbot to exhibit a form of intelligence, that is not actually the case. At a high level of abstraction, all that is happening is that the chatbot is drawing upon its stored data to produce output that will contain words from that data arranged according to patterns found in the data; it is not actually thinking or creating anything new. The program simply re-expresses information (words and lexical patterns) already encoded within its training dataset, based upon the input fed to it.

Obvious limitations of a statistically based rearrangement of the training dataset

The third aspect of AI is that the uses and pitfalls of chatbots for barristers become relatively obvious once one understands how they work. The most important are set out below.

First, because a chatbot is not truly intelligent, it is not a replacement for human intelligence. In the context of the Bar, a chatbot cannot exercise judgment nor strategise. It may produce output that replicates thinking, expressed in material that it was trained on, but it will not produce new insights other than by chance. Using a chatbot to analyse a problem without applying independent thought invites disaster.

Secondly, because output is produced by drawing upon data in the training set in a probabilistic (i.e. statistical) fashion, chatbot output may combine familiar, legal-style writing with words or phrases drawn from the training set that are plucked out of context. The result is very persuasive and confident-sounding output that resembles traditional legal writing, but which contains statements that are often wrong in whole or part. This is usually called 'hallucinating'. AI hallucinations can be difficult to spot, particularly where they appear alongside correct material. As an example, when ChatGPT-4 was asked, 'What did the case of *ROKT v Commissioner of Patents* decide?' the first two paragraphs of its answer are:

> The central question in this case was whether ROKT's invention, which related to a digital advertising method and system, was a patentable invention under Australian law. ROKT argued that their invention involved an innovative solution to the problem of online advertising and should be granted a patent. On the other hand, the Commissioner of Patents argued that the invention merely involved an abstract idea implemented using generic computer technology and did not meet the requirements for patentability.

> The High Court ultimately held that ROKT's invention was patentable. They concluded that the invention offered more than a mere abstract idea and that it provided a practical application of the idea using specific computer technology. The court emphasi[s]ed the importance of considering the actual contribution made by an invention rather than focus[s]ing solely on the abstract idea behind it.

The first paragraph is entirely correct, while the second paragraph is entirely wrong. The High Court did not hold anything; it refused special leave, leaving in place the decision of the Full Federal Court, which held that the invention was ***not*** patentable. Anyone relying upon this output without checking it would be led into significant error. A stark and now infamous example of the dangers of hallucination is the unfortunate American lawyer who relied upon non-existent cases produced by ChatGPT.[8] The supposed fragments of those cases can be recognised as more or less random re-expressions of parts of real cases contained in the dataset.

Thirdly, patterns that exist in the dataset are able to be replicated in the output, and the more frequently the pattern or word occurs, the higher the likelihood that it may be reproduced. This means that if there are stereotyped, biased or offensive statements or speech patterns contained in the dataset, there is a chance that they may be replicated in output. Indeed, in 2019, a version of ChatGPT was trained on a dataset comprising posts on the Reddit social media platform; because those posts contained offensive and abusive content (including, apparently, white supremacist posts), the version was held back as 'too dangerous to release' because it produced output containing similar statements.

Fourthly, due to the way chatbots work, there is no guarantee that they will respond the same way to the same question each time. This is not just due to their probabilistic nature. Changes to the dataset, the neural network and the algorithms employed can mean that output for the same input may differ over time. Equally, questions posed to a chatbot in slightly different ways may produce very different output. In general, the more precise the question asked, the more likely is the answer to contain correct information because the output usually strives to incorporate the input. For example, asking ChatGPT 4 'What is the limitation period on a debt?' did not produce a reference to the special case of 12 years for actions on secured debts under s 42(1) of the *Limitation Act 1969* (NSW), whereas asking 'What is the limitation period on a debt in New South Wales?' did include s 42 in the answer.

Fifthly, it is increasingly common for chatbots to take input submitted to them and add it into their dataset, and the terms of use of some chatbots now explicitly allow this. The consequence is clear: any data inputted into a chatbot may be stored and then reproduced to any other user of it. This means that sensitive or confidential information – including privileged information – should never be entered into a chatbot. A number of companies have discovered this recently, to their cost.

Sixthly, the precise content of training datasets is typically kept secret, which means it will not be clear to a user what sources may be drawn on. Additionally, chatbots can also be subjected to limitations or exclusions imposed by their creator: they can be set up not to discuss certain topics, or not to include certain data sources in certain contexts. This may, or may not, be made clear to the user, in which case it may not be clear what sources are ***not*** drawn on, or what output will not be displayed. Each limitation imposes limits upon the quality and usefulness of information produced by a chatbot which may not be known to the user.

Where to from here?

As Nobel laureate Niels Bohr (perhaps apocryphally) said, prediction is very difficult, especially if it's about the future. However, it seems likely that use of AI will be a feature of the practice of law very soon, although it is likely to be in the form of specialised legal AI (trained on narrower but much deeper datasets) rather than generalist chatbots like ChatGPT (since their legal coverage is thin and swamped by irrelevant, non-legal material).

At least one major legal publisher has indicated that it will soon release a specialised legal AI, which will have some significant architectural differences from ChatGPT. Although it will be trained on solely Australian legal material (i.e. all law reports and unreported judgments, and secondary sources, probably segmented by topic and jurisdiction), it will not be a 'pre-trained' model that operates using statistics, but rather a search engine that uses AI to combine and summarise the material found using a chatbot-style search. That is likely to reduce (perhaps even avoid) the chances of a hallucination, since the system will be starting with trusted material and trimming it down, rather than generating novel text on a purely statistical basis. The system is also intended to allow summarising of cases or documents uploaded by the user (which will not be incorporated into a pre-trained model, to avoid confidentiality concerns), and its generative function will be limited to drafting letters and clauses – although one would expect more kinds of documents will be added in the future.

It is also likely that these specialised models will become important, perhaps essential, for legal research. Indeed, Google has already recognised that GPT-based AI is an existential threat to its general search business (which, so far as is known, operates according to a very different kind of algorithm), and the chatbot-style interface is certainly easier for users than the existing Boolean logic–based interfaces of major providers. It also seems clear that AI hallucination will be a pressing problem for such services (and for all fact-based services that require accuracy) and that efforts will be made to address it, although – as just noted above – this would likely require a significant change from an algorithm based purely on probability. **BN**

ENDNOTES

1 <https://fortune.com/2023/05/08/henry-kissinger-ai-nuclear-weapons-warning-risk/>.

2 For a good explanation of the design of Deep Blue, the first chess AI to beat a human grandmaster, see M Campbell, AJ Honae Jr and F Hsu, 'Deep Blue' (2002) 134(1–2) in *Artificial Intelligence* 57; available at <https://www.sciencedirect.com/science/article/pii/S0004370201001291>.

3 A somewhat technical explanation of how this can be done can be found in E David, HJ van den Herik, M Koppel and N Netanyahu, 'Genetic Algorithms for Evolving Computer Chess Programs' (2014) 18(5) *IEEE Transactions on Evolutionary Computation* 779: available at <https://arxiv.org/pdf/1711.08337.pdf>.

4 A good explanation can be found on the '3Blue1Brown' YouTube channel <https://youtu.be/aircAruvnKk>.

5 The other main source of interest is AIs that generate images. They have a number of similarities with text-generative AIs but are beyond the scope of this article.

6 This is, in fact, the basis of the Turing Test for AI, which assesses whether a human can reliably tell that a dialogue partner in this form of dialogue is a computer or another human; see <https://www.turing.org.uk/scrapbook/test.html>.

7 An interactive version is available at <http://psych.fullerton.edu/mbirnbaum/psych101/eliza.htm > and a short video of ELIZA in action and an interview with its creator is at <https://youtu.be/RMK9AphfLco>.

8 The docket is at <https://www.courtlistener.com/docket/63107798/mata-v-avianca-inc/?filed_after=&filed_before=&entry_gte=&entry_lte=&order_by=asc>.