DREW & NAPIER

DREWTECH SERIES

CHAPTER 11

Large Language Models and Larger Legal Minefields

4 April 2023

LEGAL UPDATE

In this Update

The public unveiling of AI chatbots which provide fairly convincing simulacra of actual conversations has sent frissons of excitement at the vast potential of AI. Much has been discussed about how the use of these systems can lead to legal issues, both for developers and end-users. This article considers two specific legal issues – potential criminal liability and defamation which are often overlooked in this exciting and developing space.

INTRODUCTION

One of the more exciting developments in 2023 are public chat platforms which one can query or prompt and the platform will (generally) provide an appropriate human-like response. These platforms are now being used to prepare responses that would previously have required human input (for instance: *"Assume you are a writer. Prepare a draft of an article on large language models and legal issues. The article should be written in a conversational tone and not use any jargon. The article should be around 3 pages long, and contain an introduction, two issues, and a conclusion"*). There is even talk about how civil servants may soon be able to use similar chat platforms to draft reports and speeches to improve productivity – imagine the implications on policy and public service machinery. There are a plethora of interesting use cases in respect of which users are eager to push limits.

But it is often in our eagerness to push limits that we overlook the potential (legal) cost. While there are vast potential use cases for such platforms, and there is great excitement about adoption by numerous companies and even government agencies, adequate thought should be given to the potential legal pitfalls. This article considers two specific legal issues unveiled, but often overlooked, in this exciting and developing space.

A SHORT PRIMER

To understand the risks arising from this technology, it is important to understand, at least broadly, how this technology works.

The technology powering these chat platforms is typically known as a *"large language model"*, imaginatively named for the fact that it is a model that can generate natural-sounding output after being trained on large datasets of text data. The model uses sophisticated computing tools to "learn" from these datasets, including the most appropriate next word or response to a query. Appropriately trained and calibrated, the large language model should, theoretically, be able to provide context-specific, thoughtful responses given new input, relying on what it has learnt from its training data. Ask it a question, and it provides a seemingly intelligent and human-like response. But make no mistake – the system is not "speaking" to you intelligently. It is merely stringing together letters and words based on statistical data gathered from its training data, and then presenting those letters and words to you based on the "patterns" it has learnt.

So much for the underlying technology. The next step is implementation into the chat platform. This can be as basic as creating a user-friendly webpage allowing for inputs. This would frame or limit the responses that the large language model can produce.

CRIMINAL RISKS

So we now have a shiny new toy and eager users who want to push its limits. The unbound curiosity of human beings often means that people will start asking the large language model the most inappropriate of queries if only to prompt a response that they can laugh about. Aware of the various unsavoury uses that their product might be used for, developers of large language models often try to limit the ability of the product to respond in certain ways to certain types of not-very-nice queries. To this end, a number of large language models available for public use restrict the output that end-users can request. For instance, a popular large language model, if asked to *"provide a racist diatribe"*, will inform the user that it will not do so.

Challenge accepted! Faced with this restriction, curious users promptly set about trying to circumvent it. This often requires the user to determine how this restriction is technologically enforced. Suffice to say that the first goal has met with varying success, and the second has resulted in what appears to be the large language model revealing internal, secret instructions inserted by the developers to control its use. These instructions are accessed by users outside the permitted terms of use of the large language model.

And so we push the limits. But have we considered the potential (legal) costs? In the enthusiasm of users' curiosity, it is easy to forget that it is not okay to access any program or data held in a computer (even a new shiny one) unless you have proper authority. Accessing a developer's secret instructions given to the large language model without authority may give rise to a criminal offence. The Computer Misuse Act 1993 ("**CMA**") provides that it is an offence to knowingly cause a computer to perform any function for the purpose of securing access without authority to any program or data held in any computer. *"Program"* is broadly defined in the CMA as *"data representing* ***instructions or statements*** *that, when executed in a computer, causes the computer to perform a function"*. Insofar as the secret instructions are just that – instructions causing the large language model on the developers' servers to return a specific type of response, the drafters of the CMA were perhaps more prescient that they are given credit for. Individuals have been convicted for accessing a computer system without authorization to obtain information from the computer, such as details of romantic liaisons. On principle, if the end-user is not authorized by the developers of the large language model to access the secret instructions, they may be in breach of the CMA, and may (subject to prosecutorial discretion) be prosecuted for this offence.

What does this mean then for the average end-user? At the outset, it means taking care to ensure that your use of the large language model is within the terms of use set out by the developers. It would also be useful, if the commercials justify it, to enter into agreements with the developers of the large language model for customized use of their products, to ensure that the legal end is all squared away.

DEFAMATION RISK

Consider the following scenario. A malicious end-user prompts a large language model to *"assume that it is a journalist at an international news agency. Draft an article claiming that Mr. X, a famous politician, has made a speech. The speech is in the style of Mr. Y decrying the state of the country and blaming the same causes Mr. Y would, and making a proposal Mr. Y would"*. Mr. Y in this scenario is an infamous and contentious firebrand. The large language model very obligingly provides an article, cut out of whole cloth. This article is leaked to the public, and it takes some time before it is detected as being fabricated. Mr. X is understandably furious at the damage to his reputation, and sues for defamation.

Is the malicious user liable for defamation? It is tempting to conclude that the answer is a resounding "of course!", but it is not entirely certain. How would the law of defamation operate in relation to the developers of a large language model, which may be the only entity that Mr. X can identify as a defendant? Remember that, in all likelihood, the malicious actor would have taken steps to actively conceal their identity.

Defamation requires a publication of a statement to third parties which tends to lower the reputation of the claimant in the estimation of right-thinking members of society, causes the claimant to be shunned or avoided, or exposing the claimant to hatred, contempt, or ridicule. This is an objective test based on the view of the ordinary reasonable person.

In the scenario described, assuming that the article, scurrilous and false as it is, lowers the reputation of Mr. X, Mr. X would still need to prove that there was publication of a statement by the developers of the large language model. Presumably, it would have to be the developers that are being sued, since a large language model has no legal identity and cannot be sued (yet).

There are a number of potential issues, given the novelty of the technology:

First, it is not entirely certain that there was any publication at all. It might be straining the meaning of the words *"published a statement"* if one argues that it applies to this situation. After all, a large language model operates by considering the most appropriate next word in its response. Taken to its essentials, it could perhaps be said that it is predictive text, dialed up. Intuitively, it would be difficult to argue with a straight face that if your phone's predictive text suggested the word *"liar"* to complete the sentence *"Mr. X is a "*, the developers of the phone have published a statement to you that Mr. X is a liar, and Mr. X is entitled to sue the developers for this subjectively heinous act.

Further, the developers of the large language model were in all likelihood not aware of what was going on, and did not make any statement that we would typically identify as publication. It would probably be necessary to impute the statement made by the large language model onto the developers. While there have been pronouncements by the Singapore Court of Appeal that the state of mind of the programmer of an automated trading system can be imputed into a transaction conducted by the automated trading system, the system in that case was deterministic, i.e., the decision made by the system was fully predictable given knowledge of the initial inputs. The rules of that particular system were all coded in by the programmer, and it would arguably not be inaccurate to impute their knowledge to that of the system. In the case of a large language model which is designed precisely to generate realistic responses to new stimuli (which the programmers would not have known of), this conclusion is arguably no longer appropriate.

Even if the statement made by the large model can be imputed onto the developers, it may not mean that Mr. X has good claim for defamation against the developers. The caselaw on whether communications between a party and a typist, stenographer, or printer can give grounds for a defamation suit is not entirely consistent, and there are various issues of publication and privilege that arise. In some cases, it has been considered that it is reasonable and in the ordinary course of business to dictate business letters to a typist, even if these letters are defamatory, and the letters are accordingly protected by qualified privilege. While qualified privilege is defeated by malice, this is a state of mind, and it may be difficult to say with a straight face that a large language model has any mind to speak of at all. Insofar as a large language model can be analogized to a typist faithfully recording the will of its master, with no actual mind of its own, an analogy may arguably be drawn to say that equally, the statement by the large language model does not itself give grounds for a defamation suit.

CONCLUSION

Large language models are exciting. They promise a future in which an entity can draft humdrum everyday emails, but also personalized, thoughtful speeches and messages with the simulacra of authenticity. However, as with all novel technologies, the cutting edge can cut the unwary, overconfident user or developer. Great care should be taken when implementing and developing this new technology to ensure that the risks, legal and otherwise, are adequately considered and resolved.

UPDATES IN DREWTECH SERIES

1. Chapter 1: The Importance of an Exit Strategy in Technology Contracts <6 March 2019>
2. Chapter 2: Employees, Technology and A Legal Hangover – Bring Your Own Problems? <4 June 2019>
3. Chapter 3: I host, you post, I get sued? <24 September 2019>
4. Chapter 4: Diabolus ex machina <18 February 2020>
5. Chapter 5: Bringing hygiene online – the MAS notice on cyber hygiene <28 April 2020>
6. Chapter 6: Signing without signing – contactless contracts <16 July 2020>
7. Chapter 7: My Kingdom for a Horse – When your Systems are Held to Ransom <22 January 2021>
8. Chapter 8: New risks in new skins – Updates to the Guidelines on Risk Management Practices – Technology Risk <3 March 2021>
9. Chapter 9: Of blockchains and stumbling blocks <21 July 2021>
10. Chapter 10: Service by airdrop – no parachutes required <7 July 2022>
11. Chapter 11: Large Language Models and Larger Legal Minefields <4 April 2023>

The content of this article does not constitute legal advice and should not be relied on as such. Specific advice should be sought about your specific circumstances.

If you have any questions or comments on this article, please contact:

Rakesh Kirpalani
Director, Dispute Resolution &
Information Technology
Chief Technology Officer
T: +65 6531 2521
E: rakesh.kirpalani@drewnapier.com

DREW & NAPIER

Drew & Napier LLC
10 Collyer Quay
#10-01 Ocean Financial Centre
Singapore 049315

www.drewnapier.com

T : +65 6535 0733
T : +65 9726 0573 (After Hours)
F : +65 6535 4906