

Liability for defamation by AI

1 November 2023 7 min read

Generative AI has exploded into the public consciousness and into widespread use with the emergence of language processing tools (or large language models (LLMs)) such as ChatGPT. The objective is to mimic human-generated content so precisely that the artificially generated content can be indistinguishable.

This is achieved by assimilating and analysing original content on which the tool has been trained, as supplemented by further learnings from prompts and its own generated content – essentially by learning patterns and relationships between words and phrases in natural language to repetitively predict the likeliest next word in a string based on what it has already seen and continues these predictions until its answer is complete[1].

A curious feature of LLMs is that they sometimes produce false and even damaging output. Instances of lawyers including fictitious AI-generated case law in their submissions to court are already well known, but LLMs can and do go further.

Outputs can generate false and defamatory content that has the potential to cause a person actual reputational damage. This can even include fabricating non-existent "quotes" purportedly from newspaper articles and the like.

This tendency to make things up is referred to as hallucination, and some experts regard it as a problem inherent in the mismatch between the way generative AI functions and the uses to which it is put. For the time being, at least, it is a persistent feature of generative AI[2].

This inevitably raises the question of where legal liability rests when LLMs generate false and harmful content.

In the USA, much of the debate has centred around whether the creator of the LLM – such as OpenAI in the case of ChatGPT – can be held liable in light of the statutory protection afforded to the hosting of the online content of other content providers under the U.S. Code 230, although it appears that the generally held view is that Generative AI tools do not fall within the protection afforded under this law, as it generates new code and does not merely host third party content.

In the EU, the European Commission's proposed AI Liability Directive, currently still in draft form, will work in conjunction with the EU AI Act and make it easier for anyone injured by AI-related products or services to bring civil liability claims against AI developers and users[3].

The EU AI Act, also currently in draft form, proposes the regulation of the use and development of AI through the adoption of a 'risk-based' approach that imposes significant restrictions on the development and use of 'high-risk' AI.

Although the current draft of the Act does not criminalise contravention of its provisions, the Act empowers authorised bodies to impose administrative fines of up to 20,000,000 EUR or 4% of an offending company's total worldwide annual turnover, for non-compliance of a particular AI system with any requirements or obligations under the Act.[4]

In the UK, a government White Paper on AI regulation recognises the need to consider which actors should be liable, but goes on to say that it is 'too soon to make decisions about liability as it is a complex, rapidly evolving issue'[5].

The position in South Africa is governed by the common law pertaining to personality injury.

The creator of the LLM would presumably be viewed as a media defendant, meaning that a lower level of *animus iniuriandi* – namely negligence – would be required to establish a defamation claim than if the defendant were a private individual. What would constitute negligence in the case of a creator of an LLM that is known to hallucinate is an open question, which may depend on whether reasonable measures to eliminate or mitigate the known risks could have been put in place by the creator[6].

What is clear is that disclaimers stressing the risk that the output of the LLM will contain errors – which AI programmes often contain – would not immunise AI owners from liability, because they could at most operate as between the AI company and the user, but would not bind the defamed person.

But on a practical level, the potential liability of the AI creator would be of less importance to a South African plaintiff, because the creator would have to be sued in the jurisdiction where it is located (except in the unlikely event that it had assets in SA capable of attachment to found jurisdiction), rendering such claims prohibitive.

The potential liability of the user of the LLM, who then republishes the defamatory AI-generated output, is another matter.

Firstly, it is no defence to a defamation action to say that you were merely repeating someone else's statement. Secondly, the level of *animus iniuriandi* required would depend on the identity of the defendant.

If the defendant was a media company – for example, an entity that uses AI to aggregate and summarise news content – then only negligence would be required, and that might consist of relying on an LLM known to hallucinate without putting the necessary steps in place to catch false and harmful output.

If on the other hand the defendant was a private individual using the AI to generate text, then the usual standard of intent would apply, which would obviously make a claim much harder to establish. Intent however includes recklessness.

It remains to be seen whether our Courts would consider it reckless to repeat a defamatory AI-generated statement in light of the caveats that AI creators have published against the use of their AI tools.

For example, OpenAI has provided users with a number of warnings that ChatGPT "can occasionally produce incorrect answers" and "may also occasionally produce harmful instructions or biased content".[7]

It remains to be seen what approach the Courts will adopt regarding false and defamatory AI-generated content.

We anticipate that in dealing with these questions, the Courts will have to engage with questions of public policy, such as balancing the competing interests of reputational rights with not imposing undue burdens on innovation and the use of new technologies.

As LLMs are increasingly integrated into larger platforms (e.g. search engines), so their content will be published more widely and the risk of reputational harm to individuals referred to will increase[8].

This area of delictual and product-related liability can be expected to develop rapidly in coming while.

Footnotes

[1] Natasha Lomas, "Who's liable for AI-generated lies?", 1 June 2022

[2] Matt o' Brien, "Tech experts are starting to doubt that ChatGPT and A.I. 'hallucinations' will ever go away: 'This isn't fixable'", 1 August 2023

[3] Womble Bond Dickinson, "AI and liability 2023: a guide to liability rules for Artificial Intelligence in the UK and EU", 9 May 2023. The Directive will *inter alia* include a presumption of causality, i.e. a rebuttable presumption that the action or output of the AI *was* caused by the AI developer or user against whom the claim is filed.

[4] [Regulation of The European Parliament and of The Council](#)

[6] Volokh *op. cit.* discusses the following possible precautions which may or may not be technically feasible: quote-checking; avoiding quotes altogether; double-checking output; notice-and-blocking mechanisms for reported falsehoods; and discontinuing earlier versions of the LLM when new versions prove materially more reliable. The last-mentioned is particularly interesting, because leaving an earlier and admittedly less reliable version in the market, and making it free to use, could reasonably be construed as negligent in and of itself.

[7] [What is ChatGPT?](#)

[8] Farrer & Co, "Can ChatGPT and other generative AI tools be liable for producing inaccurate content?", *per* Ian de Freitas and Thomas Rudkin, 6 July 2023



Preeta Bhagattjee

Head of Technology & Innovation

[View Profile](#)