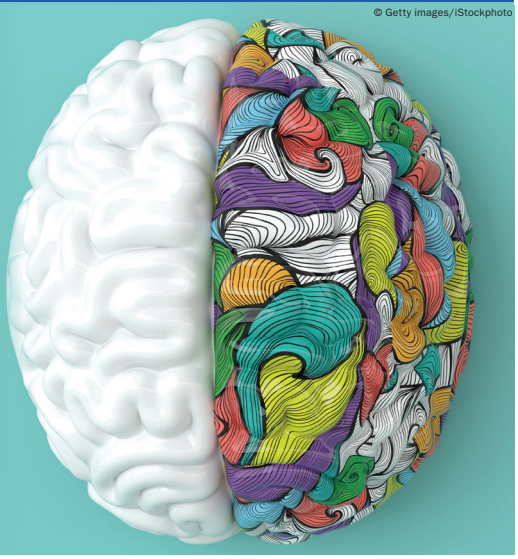

AI hallucinations: fact vs fiction

Artificial intelligence tools are not (yet) above creating false information: who could be liable for the serious harm suffered as a result of publishing that information? Chloe Flascher examines a thorny legal issue

IN BRIEF

▶ While we await specifics on AI regulation in the UK, this article examines the libel risk in this jurisdiction faced by users of generative AI systems who republish false output data.

▶ It examines the importance of companies devising internal policies on the use of generative AI in the workplace which properly factor in the risks faced by users republishing false output data about third parties.

It is well-publicised that ChatGPT recently invented a sexual harassment scandal, naming a real law professor as the accused (citing a fake *Washington Post* article as evidence in support of the allegation). Not only did no such article exist, but the real professor had never been accused of harassing a student, nor had he been present on the trip to Alaska described by the chatbot during which the purported sexual harassment took place. In June, a US lawyer facing potential sanctions was provided an opportunity to explain how it came about that he found himself submitting a court document which relied on non-existent judicial opinions and citations—all generated by ChatGPT. The lawyer said that he simply 'did not comprehend that ChatGPT could fabricate cases'.

While we await detailed outline from the UK government as to how its proposed regulatory framework will work in practice, the reality is that employees *do* continue to use generative artificial intelligence (AI) tools such as ChatGPT for work purposes, whether that be with or without employer knowledge. For anyone reading this article who has never used generative AI, when you log in as a user of the latest version of ChatGPT, it pops up with a message which states that while there are 'safeguards in place, the system may occasionally generate incorrect or misleading information and produce offensive or biased content'. It also states that 'it is not intended to give advice', and it 'may produce inaccurate information about people, places, or facts'. The system also says that it has 'limited knowledge of world events after 2021'. What is clear is that in its desperate attempt to answer questions asked of it, generative AI systems such as ChatGPT are (at the very least) spreading misinformation and (at the very worst) generating libellous content.

If you are the individual whom the false information is about, it will be very obviously untrue. But what if you are the employer of the employee using ChatGPT and unknowingly republish the false and defamatory statement contained within its answer? What if you are the employee unknowingly citing sources which don't exist in support of those false allegations because the generative AI system has hallucinated in its attempt to impress you with its encyclopaedic knowledge?

What is hallucinated output?

To hallucinate is generally defined as the act of seeming to see, hear, feel, or smell something that does not exist. In the field of generative AI, it is understood to describe instances where generative AI models create content that either contradicts the source or creates factually incorrect outputs under the appearance of fact.

Academics have voiced that it is really not at all surprising that generative AI tools 'hallucinate' in circumstances where the technology is built to generate text and language, as opposed to be factually accurate. In other words, generative AI systems such as ChatGPT are created with the goal of generating new text based on replicating existing language, rather than in order to answer questions or be factually correct.

What is the problem?

Generating language and expecting that language to be factually correct are two very different goals. While the goal of generative AI is to replicate language, as humans we don't simply 'speak'—we actually look up information, recall information, and deduce information from our lived experiences. This value that we add as humans cannot (yet) be replicated by generative AI, meaning that in its attempts to impress its master and 'fill in the gaps', false statements are being produced.

If a false and defamatory statement created by generative AI is published and is likely to cause or has caused serious harm to subject of the statement, this begs the question: who could be liable under English defamation laws for the serious harm suffered as a result of that publication?

In theory, the website company controlling or hosting the service could be liable (applying the English case of *Godfrey v Demon Internet Ltd* [2001] QB 201). In reality, however, given that most entities which develop and host generative AI products such as ChatGPT (OpenAI) and Google Bard (Google LLC) are not domiciled in England and Wales but in the US, it would not be possible to bring a libel claim against a US entity in this jurisdiction without tackling the main jurisdictional hurdle faced by mounting an English libel claim against a US-based publisher: ie by showing that of all the places where the defamatory statement has been published, England and Wales is clearly the most appropriate forum.

What if I am only repeating the statement?

In practice, the higher legal risk is that faced by *users* of generative AI systems such as ChatGPT who are *repeating* the false and defamatory statements produced by the system about an individual or entity. That user or their employer or publisher (if different) could also be liable for its repetition. The output data containing the libellous content may be, for example, published in a news article by the user or employer of the user, or in a research paper, or contained in an email to a third party.

Under a well-established rule known as the so-called 'repetition rule' the

republication of a false and defamatory statement amounts to a new publication for the purposes of English libel law. In other words, the user cannot escape liability for defamation by attributing a statement to another person, and it will be no defence to say that it was simply a republication of a false statement produced by generative AI.

Could publication be defended?

This article deals with AI hallucinations and so it is naturally assumed that the statement published is false, rather than simply inaccurate. This means that a defence of truth would not be available, as it would in those circumstances not be possible to prove that the 'sting' of the libel was substantially true. Subject to the extent to which pre-publication verification processes were made by the user of the AI system prior to repeating the statement and depending on the context in which it is published, it may be that its repetition is defensible under s 4 of the Defamation Act 2013 (DA 2013). Section 4, DA 2013 requires that the defendant reasonably believed that publishing the statement was in the public interest. This is both an objective test (that the statement was on a matter of public interest) and a subjective test (that this was 'reasonably believed' by the defendant).

The question of whether or not a public interest defence is available to any defendant is incredibly fact-specific; however the Supreme Court decision of *Serafin v Malkiewicz and others* [2020] UKSC 23, [2020] All ER (D) 13 (Jun) (applying the Court of Appeal decision of *Economou v de Freitas* [2018] EWCA Civ 2591) makes clear that the courts determine a belief to be reasonable for the purposes of a public interest defence only if it is 'arrived at after conducting such enquiries and checks as it is reasonable to expect of the particular defendant in all the circumstances of the case'.

The AI chatbot providing the user with the false statement certainly will not have considered whether or not onward publication would be a matter of public interest. In defending the republication of the defamatory statement, this means that it will be fundamental to understand what checks have been made (in context of what it is 'reasonable to expect' of them) prior to onward republication of the false statement.

It is also a defence to an action for defamation for the defendant to show that the statement was an honestly held opinion under s 3, DA 2013. The defence is defeated if the claimant shows that the defendant did not in fact hold the opinion. The defendant would have to prove that the statement was one of opinion (rather than opinion posed as fact) and that the user of the generative AI honestly held the opinion (despite having relied on AI-generated output).

These are difficult legal questions which will be tricky to resolve, but it is clear that republication would not be defensible as on a matter of public interest or honest opinion in circumstances where a user of generative AI *unknowingly* republished the false statement.

It is worth noting that libel risk is not the only legal risk here. To repeat false statements about third parties produced by generative AI may expose the repeater to liability under the tort of misuse of private information. It is a well-established principle under English law (as derived from the leading case of *McKennitt v Ash* [2006] EWCA Civ 1714) that a claim for misuse of private information can be brought in relation to information which purports to be private information about the claimant, whether it is true or false.

Risky business

If you work in a regulated profession in the UK, then it will be obvious that it is unwise and potentially unethical and improper to rely heavily on unverified statements generated from an AI chatbot as the sole source of research or as part of the provision of advice or a document being sent to a third party, such as the court, without carrying out a further significant (human-led) review of that output from the AI chatbot. This may be much less obvious to other individuals working in unregulated professions who (whether out of intrigue, habit, or otherwise) may be starting to rely on generative AI in day-to-day workplace tasks, such as the drafting of emails or in generating the content of a sales pitch without thinking about the consequences of republishing its output unrefined.

There are employees who will be told to steer clear completely of using generative AI for work purposes. There will be employees who will be permitted to use it for certain tasks, but not for others. Among other issues which employers face when determining the adequacy of internal policies on the use of generative AI in the workplace—including the thorny one of ensuring that employees do not input confidential client information to generative AI tools—it will be important for companies to be clear on the extent of human-led verification and fact-checking they require of employees before relying on sources or statements produced by generative AI.

Matters may be complicated by the reality that it cannot be assumed that any libellous content produced by a 'hallucination' is necessarily detected. In practical terms, companies need to be alive to the issue by ensuring that their internal policies on the use of AI generative systems for work purposes are clear on what type of use they permit, and in particular on the subsequent methods and standards of verification and fact-checking they require, before any material derived from or produced by generative AI is repeated and published. **NLJ**

Chloe Flascher, associate, media & reputation law, at Withers LLP (*www.withersworldwide.com*)