5 Days, 5 Disputes

11 August 2025

Day 1

by Paul Schwartfeger

When AI Lies: Hallucinations and Falsehoods

When a generative AI (GenAI) system responds to your query with confidence, it feels authoritative, as if the machine "knows". It doesn't. Large language models (LLMs) like ChatGPT generate text by mimicking plausible sequences of words from statistical patterns in their training data, not by retrieving verified facts. They may invent details entirely, a phenomenon known as "hallucination".

Hallucinations are no longer just a curiosity for tech law blogs. In recent months, lawyers, journalists and businesses have all reported AI tools inventing allegations, misquoting sources or wrongly attributing actions to named individuals. For example, a claim was recently brought against OpenAI by a Norwegian man after ChatGPT falsely reported he had murdered his sons. Similarly, in the US, OpenAI was sued for wrongly linking a political commentator to extremist activity. These cases show the immediate reputational and legal peril posed by AI hallucinations.

Applicable legal principles

If hallucinations concern an identifiable person, then a claim for defamation under the Defamation Act 2013 may be possible. The claimant must show the words referred to them, were published to a third party and caused or were likely to cause serious harm. In an AI context, "publication" may be satisfied when the output is displayed, emailed or otherwise made accessible. GenAI can complicate the analysis because outputs are shaped by a user's prompts and may differ with each interaction.

Where false information is produced in a professional or quasi-professional context implying reliability, a claim for negligent misstatement may be arguable (*Hedley Byrne & Co Ltd v Heller & Partners Ltd* [1964] AC 465), provided a duty of care can be established. With AI, questions of duty often turn on where responsibility lies: with the developer who designed and trained the tool, the party that integrated it for a specific use, or both.

If the statement induces someone to act, such as entering into or avoiding a contract, misrepresentation may arise. However, in an AI setting, identifying the "maker" of the representation can be more complex. Was it the developer, the system integrator, or perhaps even the user who framed the prompts and shaped the system's answer? It might be all three.

A claim for malicious falsehood is also a possibility where a statement is false, published maliciously and causes financial loss. Malice may be inferred if a defendant knew, or was reckless as to whether, their AI tool could generate false information.

Potential defendants

Multiple defendants need considering, as responsibility may lie with the model provider, the organisation deploying it, or even a user if they repeat false outputs.

Evidence and interim remedies

False content can spread quickly. Interim injunctions are available, though granted cautiously in speech cases. Norwich Pharmacal orders can be used to obtain prompts and output logs from non-parties. For parties, these should be sought via disclosure. An order under CPR Part 25 may be important to preserve property, including volatile records such as cached outputs and logs. CPR rule 25.1(1)(c)(ii) includes the power to order inspection of a database, if necessary and proportionate (*Patel v Unite* [2012] EWHC 92 (QB)).

Potential defences

In defamation, defendants may argue truth, honest opinion or public interest, though these are more difficult to sustain when the "speaker" is non-sentient and generating apparent facts without sources.

In a claim for negligent misstatement, the absence of a duty, disputes about accuracy and whether reliance on an AI output was reasonable may be argued.

For misrepresentation, a specific false statement needs to be shown to have induced the act in question. The adaptive nature of GenAI outputs can complicate this. Models do not store "facts" in a database; identical prompts may yield different answers across runs, and outputs can be shaped by earlier interactions.

Remedies

Remedies include damages, injunctions to prevent repetition and statements in open court. Awards may be significant where economic loss can be quantified.

Conclusions

Hallucinations present novel evidential and attribution challenges, yet remain actionable under established legal principles. In this fast-moving arena, success will favour lawyers adept at interrogating both code and case law.

5 Days, 5 Disputes

Inspired by the release of OpenAI's GPT-5 and the rapid evolution of tools like it, *5 Days, 5 Disputes* highlights five types of legal dispute where artificial intelligence is testing established legal principles, offering insights for those handling AI claims.

Paul Schwartfeger

Paul is a barrister with 36 Stone, specialising in commercial litigation and international arbitration with a particular focus on legal disputes involving data and technology.

psc@36stone.co.uk
+44 (0) 20 7440 6900