Gradient descent into chaos – Hallucinations and inadvertent waiver arising from the use of AI

June 30, 2026

RELATED PRACTICES International Arbitration, Litigation

Regardless of the cognitive[1] and environmental[2] concerns arising from humanity's increasing use of AI which resulted recently in Pope Leo XIV issuing his first encyclical ("*On Safeguarding the Human Person in the Time of Artificial Intelligence*"), it is likely that the technology is here to stay. Legal work will be affected as much as anything else. Recent decisions from the English courts offer the opportunity to consider the implications arising from the risks of inaccuracies in the products of AI Platforms, as well as issues in ensuring that communications attract appropriate privilege and/or that parties avoid waiving privilege in materials.

Hallucinations

Despite AI being an initialism for "*Artificial Intelligence*", the AI systems and platforms are not themselves intelligent. The platforms with which most users are familiar, Large Language Models, are built on a neural-network architecture called a transformer.[3] Training involves a computationally and energy intensive process, feeding tokenised text through the network and using an algorithm called gradient descent to adjust billions of internal parameters within the transformer so that the model becomes better at predicting the next token in a sequence. The resulting neural nets take an input (either from a user, or another piece of software) and process this using the transformer to create a human readable output. The process, however, is essentially probabilistic – the Large Language Model seeks to find the text which would best follow the user's prompt.

Through limitations in training data and tokenisation, as well as misaligned training incentives (for example where models are incentivised in training to provide *an* answer, rather than just express epistemological uncertainty), the output of Large Language Models can often be incorrect. In June 2024, users could ask Claude 3.5 Sonnet, a Large Language Model produced by Anthropic following a training exercise which cost "*a few $10M's*"[4], how many "Rs" the word strawberry had. The model would confidently and unflinchingly respond:

> "*There are 2 "r"s in "strawberry". The word "strawberry" is spelled S-T-R-A-W-B-E-R-R-Y. The first "r" appears after the initial "st" at the beginning of the word. The second "r" is part of the "-rry" ending.*"

We can be fairly confident that this particular error, or "*hallucination*" is a result of the tokenising process – Anthropic's tokenisation breaks up the word strawberry into two tokens, considered separately and not aggregated, as follows:

> *straw* *berry*

Newer "reasoning" models (e.g. OpenAI's o-series, Claude with extended thinking, and DeepSeek-R1) often answer the strawberry question correctly because they are trained to produce intermediate reasoning steps that can break up the word. The underlying tokenisation limitation has not changed.

The actual machinations within the training process underlying the creation of Large Language Models are, however, little understood in practice (although there has been limited progress in interpretability research). As a result of the immense volumes of data ingested and computational manipulation involved in gradient descent, the transformers created (i.e. the algorithm which actually takes a user's input and spits out an output) are essentially a black box. It is all but impossible for computer scientists to understand *why* a model says what it does, or pin down misleading outputs to a particular section of the transformer. That leaves us in the unenviable and unattractive position, particularly for lawyers, of being unable to explain how we reach the conclusion contended for.

Although generative AI is a relatively new technology (at least in its more mature form), the courts are now replete with examples of AI drafted submissions being presented to the court and containing either misleading, or entirely false, arguments and authorities.

This is obviously an issue for laypersons, who frequently have no (or very little) background knowledge to sense check the tool's output. In *Harber* v *HMRC* [2023] UKFTT 1007 (TC), Ms Harber (a litigant in person) put forward 9 decisions of the first-tier tribunal as demonstrating that the Tribunal would find a reasonable excuse to pay taxes "*because of her mental health condition and/or because it was reasonable for her to be ignorant of the law*". The only problem was that none of the cases existed or, if they did, were not authorities for the submissions set out. The Tribunal was unimpressed:

> "*It causes the Tribunal and HMRC to waste time and public money, and this reduces the resources available to progress the cases of other court users who are waiting for their appeals to be determined. As Judge Kastel said, the practice also "promotes cynicism" about judicial precedents, and this is important, because the use of precedent is "a cornerstone of our legal system" and "an indispensable foundation upon which to decide what is the law and its application to individual cases*""

The dangers are, however, not limited to non-legally qualified users. Pinsent Masons is an international law firm, headquartered in London, with 490 partners across 29 offices and an annual revenue of £680 million in 2024/2025. Newly qualified solicitors in the London office are paid £105,000. One can safely assume that hourly charge out rates are sufficiently calibrated. However, in *Malcolm Cork & Anor* v *Smith* [2026] EWHC 1199 (Ch) Mullen J castigated the process which had led to a junior lawyer (identified only as "Lawyer A") to represent that the English Insolvency Rules allowed the Court to release a liquidator from liability without application to the Secretary of State. Such a power did not exist bit was 'hallucinated' by AI.

The judgment contains a concerning insight into how AI Platforms can be misused. The Court had ordered Pinsent Masons to file a witness statement, explaining how the incorrect submissions were made. The firm exhibited transcripts of Lawyer A's interactions with the firm's AI tool. These demonstrated that Lawyer A had unquestioningly adopted the AI Platform's analysis and failed to check the statutory provisions, despite having been warned by the AI tool that:

> "*I want to be candid with you — **I am not fully confident that I am reproducing the exact statutory wording of Rule 12.37(5)** with complete precision. The substance of the provision is as I have described in our earlier discussion, but for a submission to the court **you should verify the exact wording** against the current version of the Insolvency (England and Wales) Rules 2016 as published on legislation.gov.uk before relying on it. The last thing you want is to cite a provision to the court with inaccurate wording*."

After the Court flagged the apparent inaccuracy, Lawyer A continued to use the AI tool in the firm's response to the Court's query as to how the submissions were made. After the AI tool was provided with a full copy of the relevant rules, it confirmed that the power Lawyer A had previously been arguing for did not exist. After further prompts, the AI tool created a draft letter to the court including an "*unreserved apology*". Perhaps most concerningly, Lawyer A's response was:

> "*I **don't think we should apologise** – no"*.

While it goes without saying, in-house lawyers should therefore ensure that they properly check and independently consider any advice or representations produced by AI Platforms. Policies should be created, and effectively communicated to other employees and stakeholders, explaining that AI is not to be used for legal advice without independent, qualified, lawyer review.

Privilege issues

In English litigation and arbitration, parties to disputes will normally be ordered to give disclosure of documents relevant to the case to the other side. English law, however recognises that "*a client should be able to obtain legal advice in confidence … a man must be able to consult his lawyer in confidence, since otherwise he might hold back half the truth. The client must be sure that what he tells his lawyer in confidence will never be revealed without his consent*"[5].

As such, a party will not need to provide privileged documents to their opponents. A sufficient description (for now) of when legal advice privilege will apply is where there is:

i. A communication made in confidence
ii. Between lawyers and their client
iii. For the purpose of giving legal advice

In house use of AI for legal work – can interactions with an AI model attract privilege?

Subject to issues concerning the confidentiality of certain AI Platforms (discussed further below), a communication with an AI platform would seem to be at least capable of being confidential. It would also seem at least possible that, subject to the consideration of individual messages, they could be made for the purpose of the giving of legal advice (if the hallucination cases show nothing else, it is that members of the public are increasingly relying on AI Platforms for the provision of legal advice). It is notable, however, that many public facing platforms expressly disclaim providing legal advice. Anthropic, for example, define "*Use cases related to legal interpretation, legal guidance, or decisions with legal implications*" as high risk. They expressly require that "*a qualified professional in that field must review the content or decision prior to dissemination or finalization*".

What seems challenging, in any event, is arguing that an AI platform is a "*lawyer*". As Lord Taylor explained in *R* v *Derby* "*In Wilson v. Rastall (1792) 4 Term Rep. 753, it was decided that the privilege was confined to the three cases of counsel, solicitor and attorney*". In *R* v *Special Commissioners* [2013] 2 AC 185, the Supreme Court rejected extending the doctrine to cover legal advice on tax law from accountants.

An AI platform cannot be a lawyer. A solicitor or a barrister must be a natural person, which an AI platform obviously is not. As such, it seems unlikely that this could be satisfied. While not an English law decision, this was the conclusion reached by the US District Court, Southern District of New York in *US* v *Heppner*: "*Heppner does not, and indeed could not, maintain that Claude is an attorney*".

A potential side channel to seek to evade this restriction could be English law's acceptance of foreign lawyers for the purpose of the test, without considering their qualifications or ethical rules. As Lord Neuberger recognised in *R* v *Special Commissioners* [46] "*the court has approved* [the application of legal advice privilege] *to all foreign lawyers, without (it would seem) regard to their particular national standards, regulations or rules with regard to privilege*". One could foresee a country striving to be particularly "*AI Forward*" and approving the admission of an AI Platform as a regulated lawyer (presumably behind some sort of corporate personality). Whether English comity would extend to recognition of an AI lawyer in those circumstances remains to be seen.

It may be that an easier time could be had arguing for litigation privilege to apply. There, the main issue would seem to be whether an AI Platform could be considered to be a "*third party*". Presumably, where an AI Platform was remotely hosted, the argument in favour would rely on the imputation of the corporate personality of the AI Platform's owner.

Businesses should therefore ensure that decision makers and commercial stakeholders are aware of the need to avoid directing legal queries to AI Platforms. There is every likelihood that queries to such platforms will not attract privilege, and that (especially if litigation was not in contemplation at the relevant time) they would subsequently be producible if a dispute later arose.

What if privileged materials are put into an AI Platform?

But what if a clearly privileged document (such as an advice on the merits from counsel) is uploaded onto an AI Platform? The risks in this can be seen in the recent decision of the Immigration and Asylum Tribunal in *Munir, R (On the Application Of)* v *Secretary of State for the Home Department* [2026] UKUT 81 (IAC). This was a further case in which the use of AI led to false authorities being placed before the Tribunal. When the Tribunal sought to understand how these cases had been put forward, it learned that solicitors had used ChatGPT to refine advice emails.

The Tribunal's view of this was unequivocal:

> "*We also observe that to put client letters and decision letters from the Home Office into an open source AI tool, such as ChatGPT, is to place this information on the internet in the public domain, and thus to breach client confidentiality and waive legal privilege, and thus any regulated legal professional or firm that does so would, in addition to needing to bring this to the attention of their regulator, be advised to consult with the Information Commissioner's Office*."

The statements as to client confidentiality and professional obligations are almost undoubtedly correct. However, is it right that putting materials "*into an open source AI tool … is to place this information on the internet in the public domain, and thus to … waive legal privilege*"?

The first issue is the reference to "*open source*". Open source does not mean "*publicly accessible*". It means that the source code (the human readable version of the instructions to be converted, or interpreted, into machine code by a computer) is public. The vast majority of public websites are hosted on servers running a Linux operating system, one of the most successful examples of an open source project. That obviously does not mean that any information processed by those websites is, inexorably, public. Ironically, open source models are generally those capable of being run locally (i.e. on a user's own hardware) and therefore the *least* public. While not a matter of great consequence, it demonstrates the risks of Judges drawing bright lines in areas that they do not yet fully understand.

The second issue is whether putting materials into ChatGPT was "… *to place this information on the internet in the public domain*". There are, conceivably, two ways by which this could be said to have been effected.

Firstly, the very act of inputting the materials into ChatGPT. This would presumably be because the terms of service provide that OpenAI (the creators of ChatGPT) may "*use Content to provide, maintain, develop, and improve our Services, comply with applicable law, enforce our terms and policies, and keep our Services safe*". However, it seems challenging (at the least) to say that the hypothetical possibility of personnel at OpenAI viewing a chat entails putting information "*in the public domain*". In *Wee Shuo Woon* v *HT S.R.L*. [2017] SGCA 23, the Singaporean Court of Appeal considered whether privilege in emails had been lost by their being uploaded by a hacker onto Wikileaks. It was held that:

> "*36. … It is important to focus not only on the extent to which the information in question has become accessible but also on the extent to which it has in fact been accessed by the general public …* ***Potential, abstract accessibility is vastly different from access in fact****.*
>
> *37. Accordingly, the circumstances of each case must be examined. Consideration must be given to such factors as the likelihood of the information being accessed by the public, the degree to which the information has in fact been accessed and the extent to which the information may be appreciated and/or understood only with the specialised skills or expertise of the party seeking to make use of the information.* ***Merely making confidential information technically available to the public at large does not necessarily destroy its confidential character*** *…*"

At least from first impressions, it seems that provision to ChatGPT is in the same category – providing "*Potential, abstract accessibility*" rather than "*access in fact*". Of course, in *Munir*, the materials had been deliberately provided to OpenAI, rather than exfiltrated by a hacker – but it is hard to see why that should affect the analysis.

It is also worth noting that most free email providers (such as Gmail or Outlook) include in their terms equivalent provisions that user data can be used for the same or similar purposes. For example, Google's privacy terms allow it to use user data:

> "… *for the limited purpose of:*
>
> > *operating and improving the services, which means allowing the services to work as designed and creating new features and functionalities. This includes using automated systems and algorithms to analyse your content:*
> >
> > *…*
> >
> > *developing new technologies and services for Google consistent with these terms*"

The Tribunal's analysis would therefore seemingly mean that any privilege in legal advice, or litigation materials, sent to Gmail or Outlook email addresses would be instantly waived and their materials disclosable. That would be likely to come as an unwelcome surprise to ordinary consumers, who are unlikely to invest in an enterprise email service to engage a lawyer to advise on their claim against a plumber, or assist with their divorce proceedings. The absurd implications suggest that the analysis is unlikely to be correct.

Secondly, those proficient with AI systems can mount '*extraction attacks*' to recover fragments of the model's training data, potentially including content from earlier sessions that has been incorporated into a later training run. In support of a claim by the New York Times against AI providers, the New York Times showed that it was possible to prompt an AI model in unusual ways and obtain a verbatim regurgitation of articles used in the training process. Researchers at Stanford University were able to obtain the same results with Harry Potter and the Philosopher's Stone (albeit Americanised to the "*Sorcerer's*" stone in the article).[6]

However, again, it would seem challenging to argue that "*access in fact*" was created by the creation of a potential *vulnerability* to an extraction attack. Even enterprise grade systems, which are supposedly the right ones to be used, are vulnerable to cyberattacks.

While the Upper Tribunal's analysis seems to throw the baby out with the bathwater, the underlying argument is sound. Avoid providing confidential business secrets, including privileged materials to AI Platforms, or any other service, other than on terms which adequately preserve the confidentiality of the materials.

Conclusion

AI Platforms offer an opportunity for lawyers to perform work more efficiently, subject to adequate review of the work produced by Large Language Models. In-house lawyers, and companies, must ensure that this doesn't trickle down and create a false sense of confidence in non-legally trained employees/stakeholders, who will not be well placed to judge whether AI advice is plausible.

The application of considerations concerning privilege and AI Platforms is not settled. While it seems unlikely conversations with a chat bot will give rise to privilege, and users should be counselled to avoid giving sensitive information accordingly, the decision concerning waiver in *Munir* appears to go too far. That being said, companies and firms should also ensure that best practices are adopted and confidential/privileged information only provided to third parties on terms that confidence is adequately preserved.

---

[1] Kosmyna et al (currently only published in pre-print), *Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task*

[2] Despite ¼ of the world's population not having access to clean water, it is estimated that AI data centres will shortly consume more than 6 x more water than the sovereign nation of Denmark. *Li, Pengfei, et al. "Making ai less' thirsty'." Communications of the ACM 68.7 (2025): 54-61.*

[3] The concept of the transformer was first conceived of by Google engineers in June 2017 and permitted a huge leap beyond the recurrent neural networks which preceded them. Transformers include an "*attention mechanism*" allowing them to analyse the relationship between different tokens in a string in parallel.

[4] Per the CEO of Anthropic, Dario Amodei. https://darioamodei.com/post/on-deepseek-and-export-controls

[5] *R v Derby Magistrates' Court Ex parte B* [1996] 1 AC 487, per Lord Taylor

[6] Ahmed, A., Cooper, A. F., Koyejo, S., & Liang, P. (2026). Extracting books from production language models. arXiv preprint arXiv:2601.02671.