Title: Artificial Intelligence Hallucinations: Present and Future Impact of Generated Misinformation [Working Title]

Recipient(s): Editor of the New Law Journal

Author(s): Mark Tsagas, Catherine Hobby

Wordcount: 975/1900

Artificial Intelligence (A.I.); once dubbed the product of science fiction, has for a number of years already proven to be an impactful tool in a variety of industries and disciplines. Yet, with its manifestation in public consciousness over the last year, partially due to the prevalence of and ease-of-access to mainstream Generative A.I. Tools (GenAI), further to its potential strengths some glaring issues have also come to light. Namely, bias and hallucinations.

With a specific lens towards the latter, instances of generated misinformation that have come to be known under the moniker of 'hallucinations' can be construed as a serious cause of concern. In recent times the term itself has come to be recognised as somewhat controversial. Conceptually, it is an easy-to-understand metaphor, likening aspects of the GenAI process to the function of the human mind in its attempts to fill gaps in memory. However, this new term also serves to obfuscate the impact such 'false information' can produce, in terms of scale and liability. It is important to appreciate that these 'hallucinations' may result from a variety of different causes and are submitted to not be generated maliciously, based on existing case studies. However, although this fact is recognised, it offers little comfort to those that may be harmed when such bouts of misinformation are relied upon.

In terms of the legal profession, there has already been a recorded instance when attorneys submitted a legal brief containing multiple case references that did not exist, fabricated by an A.I. chatbot[1]. While the attorney in question espoused his innocence by stipulating that he was "unaware that its content could be false" nonetheless misinformation was relied upon, and appropriate due diligence was not conducted. Even though, the fabrications were unearthed in time, the implication that the court could have been misled subsequently culminating in a miscarriage of justice is unambiguous. In fact, in recognition of this issue Judicial Guidance was issued to English Judges in December 2023. The document discusses the key risks and issues resulting from GenAI, while offering direction on how to appropriately use A.I. with a purview towards the "Judiciary's overarching obligation to protect the integrity of the administration of justice".

In the healthcare sector, the highly publicised case of the Chatbot named Tessa[2], further evidences the dangers of A.I. hallucinations and the sincere need for oversight of potential outputs. In this instance, the National Eating Disorder Association (NEDA), an organisation centred around the express purpose of helping vulnerable individuals suffering from eating disorders, disbanded its helpline (comprised of salaried employees and volunteers) before announcing its replacement by the Tessa chatbot[3]. What makes this case of particular interest

[1] https://www.bbc.co.uk/news/world-us-canada-65735769
[2] https://www.wsj.com/articles/eating-disorder-chatbot-ai-2aecb179
[3] https://fortune.com/well/2023/05/26/national-eating-disorder-association-ai-chatbot-tessa/

is the fact that Tessa was originally and painstakingly designed to be a rules-based chatbot, void of generative elements and thusly unable to deviate from standardised pre-written responses[4]. Dr. Ellen Fitzsimmons-Craft, one of the researchers involved in the creation of Tessa, stipulated that "by design, it couldn't go off the rules" further suggesting that a rules-based design resulted from the fact that they were very cognisant of the fact that A.I. isn't particularly suitable for this particular demographic audience. Tessa was subsequently taken offline, after a very succinct period of operation. Reports suggested that it offered problematic advice that could have exacerbated the eating disorder symptoms. This ability to facilitate new responses, thus deviating from the preprogrammed answers, was submitted to be the result of the host company adding GenAI to the chatbot, as part of a "systems upgrade". This case study, considering that the subject matter relates to a person's physical and mental wellbeing, serves as a particularly concerning and poignant reminder that A.I. hallucinations can indeed have tangible consequences, especially if the recipient is ill-prepared to challenge the assertions made.

While the case studies above are particularly prominent, in effect there is no shortage of examples where hallucinations have caused some distress. Ranging from defamation of character, as in the case of Mr. Hood[5] when A.I. falsely asserted that we had been imprisoned for bribery, to claims of academic misconduct when A.I. software has reportedly incorrectly suggested its utilisation by students[6]. In response to this issue, certain industries have been able to adapt quickly and implement a variety of precautions that allow them to safeguard integrity, as evidenced above.

It is clear that this technology is already having a genuine effect on people's lives through further means that transcend the 'unintended' hallucinations. Consequently, a question could be posed; 'Since community leaders are already aware of the present issues, in the future would refinement and regulation of the technology not solve this defect in its entirety?'. While a fair question to ponder, should too much faith be put into a technological system, no matter how advanced, one need not look further than recent events to glean the potential undesirable outcome.

The airing of the TV drama Mr Bates v The Post Office in January 2024 brought public attention to the Post Office Horizon scandal of wrongful convictions on the basis of a faulty digital accounting system. The Post Office's private prosecution of innocent sub post masters using defective computer evidence is regarded as the "widest miscarriage of justice" ever seen[7]. The Horizon system was piloted in 1999 and rolled out to post office branches in 2000. The initial roll-out of Horizon was delayed by technical issues and from the start sub post masters were reporting discrepancies and shortfalls caused by faults. It was established in group litigation by 555 sub post masters in 2019 that Horizon had numerous "bug, faults and defects", and the Post Office knew that it generated false accounting shortfalls[8]. Despite this, the Post Office prosecuted sub post masters for offences of theft, fraud and false accounting and over 736 were convicted for these shortfalls. They received a court sanction,

[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10053367/
[5] https://www.bbc.co.uk/news/technology-65202597
[6] https://www.washingtonpost.com/technology/2023/05/18/texas-professor-threatened-fail-class-chatgpt-cheating/
[7] https://ccrc.gov.uk/news/the-ccrc-and-post-office-horizon-cases/
[8] https://www.judiciary.uk/wp-content/uploads/2019/12/bates-v-post-office-judgment.pdf

including immediate imprisonment for some and suffered loss of their good character, income, bankruptcy in many cases and social disgrace. This was acknowledged by the Court of Appeal in 2020 in quashing the convictions of 39 previous sub post masters[9]. The Post Office Horizon IT Inquiry was established in 2020 to investigate the implementation and failings of the Horizon system[10]. The title of the Inquiry is misleading as it is the conduct of the Post Office itself that is at the heart of this scandal, and not the faulty Horizon system. Humans and not IT are what caused the "affront to justice" found by the Court of Appeal in allowing the appeals of the sub post masters.

Evidence to the Inquiry demonstrates a complete lack of curiosity by the Post Office towards its own computer system, particularly in the actions of its own investigators. In conducting audits and investigations into shortfalls, they appear to have accepted without question the reliability of the Horizon data. They regarded it as infallible, even when contradictory accounts were provided by sub post masters. This conduct can be compared to the American attorney's unquestioning use of a A.I. chatbot discussed above. In their approach the Post investigators also failed to comply with legal obligations. A report to the Inquiry by its criminal prosecutions expert witness, Duncan Atkinson KC[11], revealed that Post Office's policies on the investigation and prosecution of sub post masters failed to comply with the Police and Criminal Evidence Act 1984 (PACE) and the Criminal Procedure and Investigations Act 1996 (CIPA), and codes of practice issued under each Act. There was a failure to comply with the duty under the CIPA Code to pursue all reasonable lines of inquiry, including those pointing away from the suspect and consider whether accounting shortfalls might lie with the computer system. This duty is of central importance to securing a human right to a fair trial, not least through achieving fair and adequate disclosure.

The Inquiry has also revealed that the Post Office were alerted to faults in the system by the sub post masters, and others within the organisation, from in its instalment. At one point the Horizon Helpline was receiving between 12,000 and 15,000 calls a month from sub post masters complaining of irregularities in the IT system. None of these led to action and concerns continued to be raised. An email from a member of the Post Office security team to the then head of Post Office private prosecutions in 2010, disclosed to the Inquiry, warned of discrepancies being detected with the Horizon IT system at 40 branches, but this did not stop the prosecutions.

Public concerns were also expressed as early as 2009 with the publication of The Computer Weekly's investigation into the Horizon system[12], but the Post Office sought to sustain an image of the robustness of the system to protect its brand. There was a culture of denial and cover-up by the senior management of the Post Office. The organisation even sacked the forensic investigating company, Second Sight that it contracted to investigate possible computer errors when it confirmed there were issues. The Post Office then spent millions defending the group litigation of the sub post masters and made a failed attempt to recuse Mr

[9] https://www.judiciary.uk/wp-content/uploads/2022/07/Hamilton-Others-v-Post-Office-judgment-230421.pdf
[10] https://www.postofficehorizoninquiry.org.uk/about-inquiry
[11] https://www.postofficehorizoninquiry.org.uk/evidence/expg0000002-duncan-atkinson-kc-expert-report-volume-1
[12] https://www.computerweekly.com/news/2240089230/Bankruptcy-prosecution-and-disrupted-livelihoods-Postmasters-tell-their-story

Justice Fraser when he found in favour of the litigants in 2019[13]. This was all part of a continued corporate projection of a falsehood now being examined by the Post Office Horizon IT Inquiry.

Unprecedented primary legislation has now been introduced to exonerate innocent sub post masters to redress the injustice caused by the Post Office's actions[14]. The company knew the consequences of defective computer evidence in the criminal trials were severe, but persisted in its actions knowing there were serious issues with the reliability of Horizon. This scandal is an example of corporate delusion in its use of IT. An utter falsehood that the Horizon system was robust was maintained by the Post Office despite knowing that it was delicate. A lack of oversight has ultimately damaged the reputation of the Post Office, possibly beyond repair with talks of a transfer of its ownership to operators[15]. More significantly it has wrecked the human lives of many sub post masters.

Overall, the advancement of Artificial Intelligence is construed to be a significant scientific breakthrough, one with a wide-reaching ripple effect that has not yet been fully realised. Although the most recent A.I. safety summit[16] is accepted to have had a positive outcome, regarding the concept of A.I. regulation, it should still be considered as an initial step. Reservations in this instance stem from the frequent use of terms such as "guardrails"[17] and "declaration"[18] both of which suggest reliance on voluntary commitments as opposed to a binding agreement. The UK's current approach is thusly not as direct as the European Union's Artificial Intelligence Act[19], which has been in development for some time. The former approach is more akin to self-regulation. To this end it is crucial that developers, regulators, and future overseers take heed of present and past cases in their attempts to create, implement and regulate A.I., generating prudent industry standards, lest we be condemned to repeat history.

[13] https://www.computerweekly.com/news/252459996/Horizon-IT-system-trial-suspended-after-Post-Office-accuses-judge-of-bias
[14] https://bills.parliament.uk/bills/3694
[15] https://www.theguardian.com/business/2024/feb/07/constructive-talks-held-over-transfer-of-post-office-ownership-to-operators
[16] https://www.aisafetysummit.gov.uk/
[17] https://www.datacenterdynamics.com/en/news/global-ai-safety-summit-kicks-off-in-uk-with-bletchley-declaration/
[18] https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023
[19] https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf