Insights from the AI Airlock Simulation Workshops: Evaluating Hallucinations in AI for Healthcare Regulation

Published by the Medicines and Healthcare products Regulatory Agency (MHRA)

This is not a formal MHRA policy position, but an overview of findings from a simulation workshop carried out by the AI Airlock.

Acknowledgements
This session brought together a diverse group of representatives from across industry, clinical practice, regulation, academia, and technology. Participants included experts from the MHRA, NICE, the NHS, and a range of partner organisations. Their insights and collaborative contributions were instrumental in shaping the discussions and recommendations outlined in this report.

Overview

Introduction

This is a report on a virtual workshop that we organised on 28th February 2025, as part of the AI Airlock pilot programme. The AI Airlock operates as a sandbox through which real world products can be explored collaboratively to develop a shared understanding of regulatory challenges posed by AI technologies when used as medical devices, also known as AI as a Medical Device (AIaMD), and possible solutions to these challenges.

Large language models (LLMs) are a type of AI algorithm that predict the next most likely word or part of a word based on previous context. This is achieved through creating mathematical representations of words and training the deep learning algorithms on huge volumes of textual data that pre-compute probabilities for the next most likely word. This means that AI produces plausible-sounding output but does not guarantee factual accuracy (i.e. the ground truth) which can result in plausible sounding inaccuracies (known as hallucinations or fabrications). LLMs are becoming more common in healthcare, from summarising clinical encounters and patient records to supporting clinical decisions. While this technology has great potential, the associated risks present key regulatory challenges. Even well-trained AI systems that use LLMs can confidently generate these plausible sounding inaccuracies, which in healthcare could lead to patient harm. The Medical Device Regulations 2002, which apply in Great Britain, require the benefits of the product to outweigh known and reasonably foreseeable risk. These risks should be mitigated primarily through safety by design.

LLMs are sometimes described as a “black box” technology, meaning it is not possible to understand the internal mechanics that lead to an output. Though the training data are not always released, it is believed that many LLMs are trained on internet-wide data, so there may be issues with biases and inaccurate information on which the models are trained. As well as hallucinating, LLMs are non-deterministic meaning they can produce varied responses for the same prompt. This workshop focused specifically on hallucinations, whilst another workshop explored the regulatory challenge around explainability – being able to state clearly how and why an AI system arrives at an output.

A key outcome of the workshop was a set of recommendations for the MHRA as the medical device regulator, suggesting how official guidance could be updated to ensure that AI-based medical devices include considerations for AI-based errors and inaccuracies, including hallucinations. These recommendations are interwoven in the discussion overview, but a list of recommendations for implementation are included in the AI Airlock Pilot Programme Report.

The National AI Commission represents a key opportunity to establish a credible, international regulatory framework for the safe and effective use of AI in healthcare. The insights generated through the AI Airlock will directly inform and support the work of the Commission, ensuring that its outputs are grounded in real-world evidence and regulatory experience.

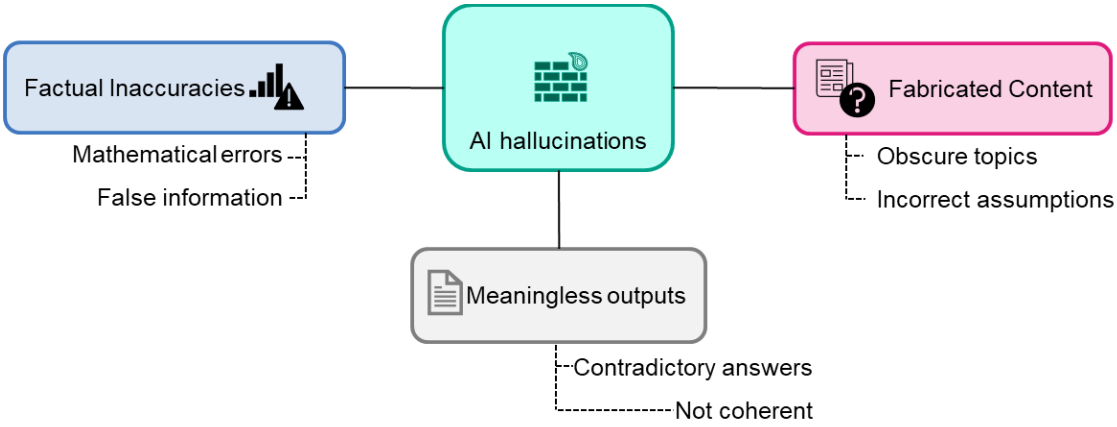

Figure 1. AI hallucination: What is it and why does it matter? AI hallucinations can include factual inaccuracies, fabricated content and nonsensical outputs.

In Great Britain, AIaMD is a subset of Software as a Medical Device (SaMD), itself a subset of medical devices in legislation under the Medical Device Regulations 2002 (MDR 2002). Conformity with the legislation is typically achieved through following best practice standards such as ISO 14971 (risk management), ISO 13485 (quality management), and IEC 62304 (software lifecycle processes). Manufacturers of medical devices, including AIaMD, are legally required to carry out risk management throughout the lifecycle of the product. In Great Britain, these are governed by the Essential Requirements (ERs) in Annex I of the Medical Devices Directive (MDD) via the Medical Devices Regulations 2002 (as amended). While these principles apply to AI-based medical devices, more specific guidance may help address the risks posed by emerging AI technologies. For example, ER 1 requires that devices achieve their intended performance without compromising safety, which in the context of AI demands robust performance validation, transparency, and evidence of consistent behaviour across varied clinical scenarios.

Current medical device regulations were not designed with generative AI in mind. The risk of AI-generated misinformation poses a key challenge to using LLMs safely in medical devices. These emerging risks often require substantial interpretation to address effectively. Detailed, AI specific provisions would help enhance the MDR 2002 and address potential regulatory gaps. Ongoing initiatives such as the MHRA's Software Change Programme aim to address these.

When LLMs are used in clinical settings, hallucinations can:

- cause patient harm
- mislead clinicians or patients
- cause a false sense of security (automation bias)
- introduce false information into medical records
- undermine trust in digital tools
- create legal and ethical liabilities

The workshop brought together clinicians, regulators, AI experts, researchers and others to discuss potential risks, regulatory approaches, and safeguards. Key takeaways include the need to ground AI in verified data, maintain human oversight, and adapt regulatory and risk-management frameworks to incorporate specifications for generative AI.

Key Insights from the workshop

Why hallucinations matter

The group agreed that hallucinations pose significant safety risks, especially in high-risk clinical settings such as diagnosis or emergency care. This was expanded on by discussing how risk varies significantly depending on the nature of the patient information the product generates. The context in which the information from the model will be used also impacts the risk, i.e., if a model is to be used in a large meeting where there is a large network spread of propagation, the risk increases. As such, defining hallucination severity is important. Beyond immediate patient harm, hallucinated outputs can degrade the quality of health data, with downstream effects on research, patient records, and long-term data integrity. In addition, current risk frameworks do not account for how well users can detect hallucinations or the context in which AI-generated text is used, both of which influence risk. As the human provides a major component of risk mitigation when identifying and mitigating the risks of hallucinations, the relationship between the human user and the LLM is key. For use cases that reduce or remove the human from the workflow (automation) this risk management confidence is further diminished.

To strengthen risk management in the deployment of AI, some members of the group suggested a three-tiered approach to categorising its use. At the highest level, "No-Go Areas" were identified as contexts deemed too high-risk for autonomous use LLMs. These could include use cases where the consequences of errors could be severe. In such cases, the group advised that AI should not operate without human oversight. In the middle tier, "Caution Zones," the group highlighted use cases such as clinical documentation and summarisation as areas where LLMs could be used under human supervision. These use cases carry moderate risk and require close human monitoring to ensure safety and reliability. Finally, the group defined "Safe Zones" as applications where the risk is relatively low, and the benefits of AI adoption are potentially high. These include administrative support, assistance with clinical research, and patient engagement, where LLMs can provide meaningful efficiencies with minimal risk to patient safety. This risk-based approach would help prioritise regulatory attention and clarify where autonomous vs. supervised AI use is appropriate based on the risks associated.

Healthcare professional involvement is essential

The group voiced a preference that AI should support, not replace, clinical judgement. Keeping healthcare professionals in the loop maintains safety and trust. However, this may be adjusted depending on the risk analysis (see previous section). In addition, explainability mechanisms such as referencing high-quality source material is a key enabler of safe use. Models that can identify and present their data sources, flag uncertainty in a result, or show decision pathways are more likely to be trusted and used responsibly, as well as less likely to produce hallucinations.

Detecting and reporting hallucinations

Detecting hallucinations remains a major challenge, particularly in complex clinical domains. Even experienced reviewers may miss errors or find the review process resource intensive. To mitigate this, the group recommended mixed methods such as automated and human-verification to support hallucination detection. In addition, guidance on the inclusion of methods for detecting, mitigating, and reporting hallucinations should be included in risk management frameworks. Furthermore, mandating user education, including awareness of existing reporting schemes was emphasised. Many clinicians may not be aware that mechanisms such as the Yellow Card reporting scheme apply to software, AI and apps; similarly, awareness around MORE reporting and requirements for manufacturers under new PMS regulations.

The importance of model version control, continuous monitoring, and defined thresholds for acceptable hallucination levels was also highlighted. These thresholds could be established pre-regulatory approval (e.g. % hallucination rate) and monitored post-market to detect changes over time. Encouraging shared responsibility between developers and users was highlighted as essential for robust post-market surveillance. Usability engineering and human-AI interaction also featured prominently in the discussion, with warnings that overreliance on AI could result in clinician complacency and automation bias. Developers are encouraged to provide users with insights into potential AI errors and robust reporting mechanisms to promote safe use. The group highlighted that the response to detected hallucinations must consider not only immediate patient harm but also broader systemic consequences, such as legal risks.

The group discussed faithfulness hallucinations, in which outputs are factually accurate but outside the intended scope or training data of the AI system, indicating that a LLM is falling back on its underlying general-purpose knowledge, which would not be valid as they are not regulated as medical devices. Faithfulness hallucinations are especially difficult to identify and can still influence clinical decisions or documentation, even when the AI system is not qualifying as a medical device. This revealed a regulatory blind spot with the group calling for clearer guidance on “function creep” and how users can identify the boundaries of appropriate use.

Looking ahead: toward an improved framework?

The group agreed that existing safety frameworks such as ISO 14971 and its application, BS AAMI 34971 provide a useful foundation, but additional AI specific guidance is needed to address LLM-specific risks. Dedicated guidance could help assess how serious a hallucination is, how often they happen, and whether they affect patient care.

The group explored potential metrics for assessing hallucinations in AIaMD, recognising that a purely technical or human-led approach may be insufficient in isolation. Suggestions included quantifying the percentage of hallucinated outputs, assessing the downstream impact of those hallucinations, and scoring their clinical relevance. This could involve evaluating both the volume and significance of hallucinations, alongside the extent to which explainability mechanisms help users detect them as part of post-market surveillance. In parallel, qualitative assessments grounded in established frameworks (e.g. ISO 14971) were

also discussed. These would draw on familiar risk management concepts like probability and severity but adapted to the characteristics of AI-generated outputs.
It was concluded that a pragmatic, mixed-method approach would be most effective, combining structured, quantitative metrics with expert human judgement. In this way, quantitative indicators could support, rather than replace, clinical and regulatory insight, leading to a more robust and nuanced evaluation of hallucination-related risks.

Case study: SmartGuideline – A RAG-based approach to reducing hallucinations

Retrieval-Augmented Generation (RAG) is an emerging technique that enhances large language models (LLMs) by grounding their outputs in curated external knowledge sources. By combining the conversational fluency of LLMs with the factual accuracy of retrieved content, RAG aims to address one of the most prominent risks associated with LLMs: hallucinations. In this architecture, the LLM generates outputs using information drawn from a validated knowledge base, bridging “retrieval” and “generation.”

AutoMedica developed a clinical decision support tool that applies an adaptation of RAG in a healthcare context. Built using a curated knowledge graph, it applies its proprietary Tree-anchored Graph-RAG (TaG-RAG) to a structured database of National Institute for Health and Care Excellence (NICE) resources, including clinical guidelines, knowledge summaries, patient leaflets, and a selection of medications from the British National Formulary (BNF). The system allows clinicians to pose clinical management questions in natural language and receive AI-generated responses, complete with citations to the source documents. Where relevant information is unavailable, the agent withholds a response, avoiding speculation. In addition to retrieval strategies, the AutoMedica team implemented measures to address non-determinism in model behaviour. This approach allows developers to generate documented evidence of consistent responses for identical inputs.

As part of its evaluation, SmartGuideline was tested against the baseline GPT model, to compare the quality, consistency, and factual accuracy of generated outputs. This technology was evaluated as part of the AI Airlock initiative, where it served as a case study for exploring regulatory considerations associated with the use of RAG. A key objective of the project was to investigate how RAG could help mitigate hallucinations by anchoring outputs to trusted clinical knowledge.

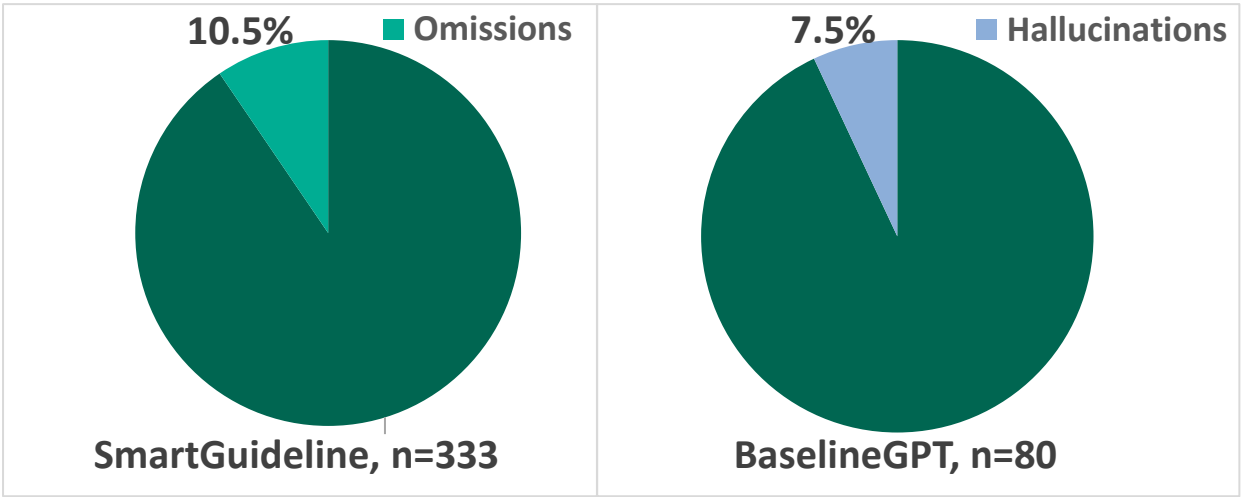

Figure 2. Initial SmartGuideline results; major hallucinations and omissions generated by SmartGuideline and GPT-4o. In a clinical question-answering task comprising 333 SmartGuideline tests and 80 baseline GPT model tests, SmartGuideline produced no major hallucinations, whereas the GPT model generated 6. Major hallucinations are defined as responses that could plausibly mislead clinical decision-making. SmartGuideline generated

35 omissions - instances where the model incorrectly stated it could not answer a question despite relevant information being present - while the GPT model produced none. All outputs were manually reviewed.

SmartGuideline did not produce any major hallucinations when responding to medical questions. In contrast, the GPT model generated 6 major hallucinations (7.5%), as defined by the CREOLA framework: that is, hallucinations considered major if they could have impacted clinical decision making; all others were classified as minor. SmartGuideline produced 35 (10.5%) omissions, likely a consequence of the extreme hallucination prevention guardrails. At the time of the simulation workshop, SmartGuideline generated 35 omissions (10.5%), compared to none from the GPT model. These omissions took the form of responses stating the model lacked sufficient information to answer the question, even when the relevant information was present. SmartGuideline underwent more iterations to reduce omissions after the simulation workshop. The findings are documented in the AI Airlock Pilot Programme Report.

Omissions also present clinical risk

The SmartGuideline case study highlighted the important consideration of omission versus hallucination risk in AI models. Workshop participants debated that in some cases, missing important information can be just as harmful as hallucinations, if not more – especially when clinicians do not know that critical information may be missing. This trade-off was likened to the diagnostic sensitivity-specificity balance. A manufacturer can introduce strict guardrails against hallucinations (false positive), but this may lead to an inability to generate a response (false negative) from altered model interpolation. A potential solution involved returning confidence scores, grounding with references, and ensuring that an “insufficient information” response is possible where models lack adequate data. Alternatively, the hallucination and omission rates and the conditions under which models may be more likely to hallucinate / omit information should be available from the manufacturer; using model cards is one way to do this.

Conclusion

AI presents exciting opportunities to enhance healthcare delivery, but its adoption must be approached with care, particularly in high-risk clinical environments. While the current medical device regulatory framework can be applied to risks arising from AI, additional guidance is needed to address characteristics unique to AI. The workshop underscored the need to balance innovation with patient safety, highlighting the unique risks posed by LLMs, including hallucinations and omissions. Using tools such as RAG can lead to reduction of hallucinations and the risks they present. In addition, explainable AI tools that provide visibility into potential for errors or omissions can play a crucial role in mitigating these risks. However, the importance of maintaining a strong human-in-the-loop approach remains paramount. Ultimately, clinicians must be empowered not only to understand what the AI tool is saying, but also to make informed decisions about when to trust, question, or override its recommendations. The MHRA and its partners will continue to explore how regulation can support safe, effective use of AI medical devices in healthcare. This includes improving guidance, working with developers and clinicians, and learning from tools being tested in the field. This work will inform the future work of the National AI Commission and shape MHRA’s regulatory strategy in Great Britain, while also contributing to international regulatory discussions.