

Stabilising Outputs? Redistributing Accountability: A Case Study of Retrieval-Augmented Clinical AI in the AI Airlock Sandbox

Jemimah Jeremiah

MSc Management of Information Systems and Digital Innovation
Department of Management
London School of Economics and Political Science

KEYWORDS	ABSTRACT
Retrieval-Augmented Generation, Clinical AI, Accountability, AI Governance, Organisational Governance, Healthcare	Generative AI systems in healthcare are often evaluated through a verification-based lens, with reliability judged by whether outputs can be traced to authoritative clinical sources. Retrieval-augmented generation (RAG) appears to address this challenge by grounding outputs in curated knowledge sources. Whether grounding resolves accountability or merely redistributes it remains unclear. Drawing on a conceptually informed case study of the AI Airlock sandbox, this paper analyses how accountability is organised when grounded clinical AI systems are developed and governed under regulatory oversight. The analysis examines how technical design, organisational governance, and regulatory oversight interact to shape accountability within grounded AI systems. It shows that grounding reduces epistemic uncertainty at the level of outputs but does not stabilise accountability within the model. Accountability instead extends across the organisational and institutional processes through which knowledge sources are selected, maintained, validated, and applied. Reliability therefore depends on continuous coordination across technical, clinical, and regulatory processes rather than on model design alone. These findings demonstrate that improving the trustworthiness of clinical AI requires governance arrangements that extend beyond technical verification to encompass ongoing organisational oversight. Grounding should consequently be understood not only as a mechanism for improving output reliability, but as a governance intervention that reshapes where and how accountability is exercised.

1. Introduction

Organisations are increasingly adopting generative AI systems (Dwivedi et al., 2023), yet their use in high-stakes settings remains constrained by concerns over reliability and accountability. Large language models can produce fluent but fabricated responses (Feuerriegel et al., 2024), meaning that correctness cannot be verified from the model alone. This condition is analogous to what Burrell

(2016) characterises as practical opacity. In healthcare contexts, this becomes a managerial problem, as decisions cannot rely on outputs whose correctness is uncertain.

Approaches to AI reliability frequently emphasise technical performance and verification mechanisms (Kroll et al., 2017; Amershi et al., 2019), overlooking how verification-based approaches reshape accountability at the organisational level

(Ananny & Crawford, 2018; Orlikowski, 2007). That framing is insufficient because it treats reliability as a property of outputs rather than of the processes through which they are produced and validated.

One response to this verification-based challenge is retrieval-augmented generation (RAG), which grounds model outputs in curated knowledge sources such as clinical guidelines. By conditioning responses on retrievable evidence, grounding can reduce hallucinations, improve traceability, and appear to address accountability by linking outputs to identifiable sources (Gao et al., 2023).

Whether grounding resolves accountability remains unclear. While verification-based perspectives suggest accountability can be stabilised through auditable outputs (Kroll et al., 2017), practice-based perspectives argue that traceability does not determine how accountability is organised in practice (Ananny & Crawford, 2018; Orlikowski, 2007). This paper argues that grounding reduces uncertainty at the level of outputs but redistributes accountability across organisational processes, challenging verification-based accounts that treat accountability as stabilised through traceability alone.

This paper adopts a conceptually informed case study approach using the AI Airlock sandbox, a regulatory sandbox for evaluating clinical AI under controlled governance conditions. The case provides an empirical setting for examining how accountability is organised when grounded AI systems are developed and governed under regulatory oversight.

2. Conceptualising grounding and accountability

Grounding refers to generating outputs by combining a large language model with retrieval from external knowledge sources (Lewis et al., 2020; Gao et al., 2023). Rather than producing answers solely from learned statistical patterns, retrieval-augmented systems condition generation on material retrieved from identifiable sources. This can make responses more traceable and auditable even when the model's internal reasoning remains opaque.

Accountability is often framed as the ability to verify and justify system outputs, rather than full transparency of model reasoning. This shifts attention from how a model produces an answer to whether outputs can be checked. In this view, accountability is achieved through verification mechanisms that demonstrate compliance with defined rules (Kroll et al., 2017), positioning accountability as a property of outputs rather than of the processes through which outputs are produced.

However, treating accountability as a property of outputs assumes that visibility is sufficient for governance. Making outputs traceable does not determine how they are interpreted in practice (Ananny & Crawford, 2018); instead, accountability emerges through the coordination of technical systems and organisational routines. This reflects practice-based and sociomaterial accounts, which treat accountability as an outcome of situated organisational processes rather than a property of technical systems (Orlikowski, 2007; Leonardi, 2011). From this view, accountability cannot be located at the level of system design alone, as it depends on how knowledge sources are selected,

maintained, and used within organisational processes.

One interpretation treats grounding as a mechanism that stabilises accountability. If responses are linked to identifiable sources, they can be verified against established knowledge and recognised standards (Kroll et al., 2017). On this view, accountability is secured at the level of outputs: what matters is not how an answer is produced, but whether it can be verified and justified.

A second interpretation suggests that grounding does not resolve accountability but redistributes it. While outputs become traceable, accountability shifts to the processes through which knowledge sources are selected, maintained, and interpreted. From this perspective, grounding relocates accountability to the infrastructures and organisational practices through which knowledge sources and outputs are governed.

The question is therefore not whether grounding improves reliability, but whether accountability can be secured through verification or must be organised through practice. This tension is examined through the case of the Airlock sandbox.

3. Case: AI airlock sandbox

The Airlock sandbox provides an empirical setting in which this tension can be examined under explicit regulatory oversight. Developed and delivered by the Medicines and Healthcare products Regulatory Agency (MHRA) in partnership with the NHS and the Department of Health and Social Care, the sandbox is a controlled environment designed to examine regulatory challenges associated with clinical AI systems (MHRA, 2025a). The

programme worked with products and prototypes at different stages of development, using simulation, virtual and controlled real-world testing to examine regulatory challenges across development and deployment.

One pilot examined AutoMedica's SmartGuideline, a clinical decision-support system using retrieval-augmented generation grounded in NICE clinical guidance. In virtual testing involving 436 medical questions, the RAG-enabled system produced no identified hallucinations, compared with 23 from the baseline model, and generated highly similar responses across repeated prompts (MHRA, 2025a). By anchoring responses in trusted clinical content, the design made outputs more readily assessable against authoritative sources. This demonstrates the verification-based expectation that grounding makes outputs auditable. Although this could be interpreted as evidence that grounding stabilises accountability, the broader governance environment suggests that accountability is not contained within outputs alone.

However, the sandbox also shows that making outputs auditable does not stabilise accountability within the model. Because responses depend on curated and approved knowledge sources, accountability shifts from verifying outputs to governing how those knowledge sources are selected, maintained, and applied. The system therefore remains embedded within a broader governance infrastructure, including NHS information governance requirements (NHS England, 2025), MHRA regulation of software as a medical device (MHRA, 2025b), and NICE requirements for evaluating clinical effectiveness (NICE, 2022). Taken together, these frameworks

establish broader obligations concerning performance monitoring, clinical validation, information governance and human oversight. Grounding adds further governance requirements because the curated knowledge sources on which outputs depend must themselves be selected, maintained and updated.

Accountability is therefore not resolved at the point where outputs become traceable. Instead, it is redistributed across organisational and institutional arrangements that govern how knowledge is produced, maintained, and used beyond the model itself.

4. Analysis: Redistributing rather than stabilising accountability

Grounding redistributes rather than stabilises accountability. By linking outputs to NICE clinical guidance, responses become traceable and can be independently assessed against recognised medical knowledge. This supports the verification-based claim that accountability can be achieved through auditable outputs rather than full interpretability (Kroll et al., 2017). Grounding therefore reduces epistemic uncertainty by allowing clinicians and evaluators to assess responses without relying solely on unverifiable model reasoning. At first glance, this appears to stabilise accountability. Yet that interpretation treats verification as sufficient for governance, even though output reliability depends on how the underlying knowledge sources are governed.

That stabilisation is therefore limited. The ability to verify outputs does not eliminate the need for judgement; it shifts where that judgement is exercised. In the Airlock

setting, accountability is enacted through processes that determine which sources are included in the knowledge base, how those sources are updated, and how outputs are interpreted in context. These processes extend beyond the model into multiple sites of decision-making, where actors govern the selection, validation, and application of knowledge. As a result, the question of whether an output is “correct” becomes inseparable from how the underlying knowledge infrastructure is governed.

This becomes evident in the broader governance environment in which the system operates. NHS guidance structures the lawful and safe use of data in AI systems through information-governance requirements (NHS England, 2025), while MHRA post-market surveillance requirements mandate continuing performance oversight (MHRA, 2025b), and NICE frameworks establish expectations for clinical evidence and effectiveness (NICE, 2022). Taken together, these requirements structure how accountability is enacted, ensuring that it remains distributed across institutional actors rather than located within the model. While these arrangements could be interpreted as reinforcing accountability through additional oversight, they also indicate that accountability is sustained through continuous organisational coordination rather than being secured through model design.

More fundamentally, accountability emerges through the interaction of technical systems, organisational routines, and institutional constraints rather than being secured within the model itself. The model’s outputs are generated through configurations of data sources, retrieval mechanisms, clinical workflows, and

regulatory requirements. Accountability is produced through the way these elements are organised and maintained in practice (Orlikowski, 2007; Leonardi, 2011). Grounding does not eliminate opacity; rather, it relocates attention from model reasoning to the governance of the conditions under which outputs are generated.

This interpretation is reinforced by the role of organisational control in structuring how outputs are used. By defining which sources are authoritative and how responses are validated, grounded systems shape what counts as acceptable knowledge and who is authorised to judge it. Drawing on Kellogg et al. (2020), this can be understood as a reorganisation of authority, in which infrastructures guide and constrain action. Accountability therefore becomes tied not only to verifying outputs but to governing the systems that define what can be verified.

Taken together, the case reframes grounding from a solution located within the model to a governance arrangement extending beyond it. The problem of model correctness becomes one of governing how knowledge is selected, validated, updated, and applied. This argument assumes that organisational governance remains effective; where it is weak or uneven, grounding may produce neither dependable outputs nor meaningful accountability.

5. Discussion and conclusion

The analysis demonstrates that the apparent technical stabilisation created by traceable outputs depends on organisational arrangements extending beyond the model. Grounding makes responses more readily assessable, but accountability remains enacted through the processes that govern

knowledge sources, validation, oversight, and use.

This implies that AI governance in high-stakes settings cannot treat accountability as a property of the model alone. Although grounding enables outputs to be assessed against recognised standards, the Airlock case shows that reliability depends on continuous practices such as monitoring, updating knowledge sources, validating outputs, and maintaining human oversight. What appears to be a technical solution therefore depends on sustained organisational coordination. Governance cannot be confined to system design but is enacted through the infrastructures and routines surrounding the system.

These claims are shaped by the conditions under which the case is observed. The analysis draws on a controlled sandbox environment in which governance structures are deliberately emphasised and closely monitored. In less regulated contexts, this redistribution may be less formalised or less effectively coordinated, particularly where oversight mechanisms are weak or unevenly applied. The focus on a clinical decision-support system limits how far the findings can be extended to domains with different forms of expertise, risk, and regulatory oversight.

Overall, retrieval-augmented generation should be understood not only as a technical approach to improving reliability, but as an intervention that reshapes where and how accountability is exercised. In high-stakes organisational settings, the challenge is therefore not simply to produce more reliable outputs, but to design and sustain the practices through which those outputs can be responsibly used.

Author contributions

Jemimah Jeremiah: Conceptualisation; Investigation; Formal Analysis; Writing – Original Draft; Writing – Review and Editing.

Data accessibility

No datasets were generated or analysed during the current study.

Ethics and consent

Ethical approval was not required because this study did not involve human participants, human material, or personal data.

Competing interests

The author has no competing interests to declare.

Funding information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

- Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). *Guidelines for Human-AI Interaction*. 1–13. <https://doi.org/10.1145/3290605.3300233>
- Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989. <https://doi.org/10.1177/1461444816676645>
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 2053951715622512. <https://doi.org/10.1177/2053951715622512>
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., ... Wright, R. (2023). Opinion Paper: “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>
- Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*, 66(1), 111–126. <https://doi.org/10.1007/s12599-023-00834-7>
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). *Retrieval-Augmented Generation for Large Language Models: A Survey*. <https://doi.org/10.48550/ARXIV.2312.10997>
- Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at Work: The New Contested Terrain of Control. *Academy of Management Annals*, 14(1), 366–410. <https://doi.org/10.5465/annals.2018.0174>
- Kroll, J. A., Huey, J., Barocas, S., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). *Accountable Algorithms*. University of Pennsylvania Law Review, 165, 633–705.

- Leonardi, P. M. (2011). When Flexible Routines Meet Flexible Technologies: Affordance, Constraint, and the Imbrication of Human and Material Agencies. *MIS Quarterly*, 35(1), 147–167.
<https://doi.org/10.2307/23043493>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
- MHRA. (2025a). *AI Airlock Sandbox Pilot Programme Report*. Medicines and Healthcare products Regulatory Agency.
https://assets.publishing.service.gov.uk/media/68ee1fb88427701993d5e02c/AI_Airlock_Sandbox_Programme_Report_Final.pdf
- MHRA. (2025b). *Software and artificial intelligence (AI) as a medical device*. GOV.UK.
<https://www.gov.uk/government/publicatio>
[ns/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial-intelligence-ai-as-a-medical-device](https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial-intelligence-ai-as-a-medical-device)
- NHS England. (2025). *Artificial Intelligence: guidance*. NHS England.
<https://transform.england.nhs.uk/information-governance/guidance/artificial-intelligence/>
- NICE. (2022). *Evidence standards framework for digital health technologies*. National Institute for Health and Care Excellence.
<https://www.nice.org.uk/what-nice-does/digital-health/evidence-standards-framework-esf-for-digital-health-technologies>
- Orlikowski, W. J. (2007). Sociomaterial Practices: Exploring Technology at Work. *Organization Studies*, 28(9), 1435–1448.
<https://doi.org/10.1177/0170840607081138>