The Paris Journal on AI & Digital Ethics, N.2, September 2026.

Regulating the Unseen: AI Accountability in UK Healthcare Sector

Mehmet B Unver[1], Iheanyichukwu Ogu[2]

[1] Hertfordshire Law School, University of Hertfordshire, UK.
[2] Tees, Esk and Wear Valley NHS Foundation Trust, UK.

Corresponding author(s). E-mail(s): m.unver@herts.ac.uk; Iheanyichukwu.ogu@nhs.net.

Iheanyichukwu Ogu contributed to Section 3.

Abstract

The integration of Artificial Intelligence (AI) into healthcare reveals persistent accountability gaps shaped by system opacity, adaptive model behaviour and distributed responsibility across the clinical and technological ecosystem. As accountability becomes a central principle of AI governance, this article develops a conceptual framework and applies it to the UK's hybrid regulatory landscape. Using the four-layer model advanced by Novelli, Taddeo, and Floridi (2024) – comprising *compliance, report, oversight* and *enforcement* – the article examines how accountability is operationalised within UK healthcare through medical device regulation, digital safety standards, transparency mechanisms and post-market monitoring. The analysis shows that although the UK has a broad assemblage of regulatory and governance tools, their fragmentation generates interconnected accountability gaps, including deficiencies in evidence generation, transparency, attribution and institutional learning. The article argues that closing these gaps requires greater *interoperability* between governance layers, clearer allocation of responsibilities across developers and deployers, strengthened institutional capability and more effective regulatory coordination. It therefore proposes a multi-layered governance model in which the four accountability layers operate through *continuous feedback loops*, enabling a lifecycle-oriented approach to the responsible development and deployment of AI in the UK healthcare system.

Keywords

Artificial Intelligence, governance, accountability, healthcare, UK.

1 Introduction

Artificial Intelligence (AI) is rapidly reshaping the delivery, organisation and regulation of healthcare within the United Kingdom (UK). From diagnostic decision-support systems and predictive analytics to clinical decision-making and workflow automation, AI promises greater efficiency, improved clinical outcomes and more effective allocation of healthcare resources. Yet the integration of increasingly complex and data-driven technologies into clinical environments also creates significant accountability challenges and associated risks (Nemitz, 2018; Wachter, Mittelstadt, and Floridi, 2017; Zerilli et al., 2019). These challenges become particularly acute as AI systems become more adaptive, autonomous and continuously learning, potentially altering their performance and behaviour after deployment. In such circumstances, accountability cannot be treated simply as an ex ante obligation to demonstrate regulatory compliance or an ex post mechanism for assigning responsibility following harm. Rather, it requires a governance architecture capable of generating evidence, maintaining transparency, allocating

DOI: 10.65701/bwyy2jqr9w

responsibility and facilitating institutional learning throughout the lifecycle of an AI system. Accordingly, accountability has increasingly been recognised as an end-to-end principle of AI governance (Jobin, Ienca, and Vayena, 2019; The Alan Turing Institute, 2023), closely connected with other foundational governance objectives, including transparency, fairness, safety and trust.

The UK currently operates a hybrid and decentralised governance model in which statutory obligations, particularly those derived from medical device regulation and healthcare law, interact with soft-law instruments, technical standards, procurement requirements, professional norms and organisational governance mechanisms. This pluralistic structure provides flexibility and can facilitate technological innovation, but it also produces fragmentation across regulatory boundaries, institutional mandates and operational stages of the AI lifecycle. This creates a particular accountability challenge in healthcare, where AI is deployed in safety-critical environments characterised by heightened vulnerability, informational asymmetries and potential consequences for patients. The central problem is consequently not simply whether appropriate regulatory instruments exist, but whether those instruments operate together as an effective and continuous accountability system.

Against this backdrop, this article examines the accountability gaps that arise within the evolving landscape of AI-enabled healthcare governance in the UK. It critically analyses the extent to which existing legal, regulatory and organisational arrangements can provide meaningful accountability across the AI lifecycle, with particular attention to the roles of National Health Service (NHS) England, the Medicines and Healthcare products Regulatory Agency (MHRA) and other relevant regulatory and institutional actors. It considers whether existing governance tools, including medical device regulation, digital clinical safety standards, procurement requirements and Algorithmic Impact Assessment (AIA) mechanisms, are capable of addressing the distinctive challenges presented by AI in clinical settings. Particular attention is given to the difficulties created by the opacity of adaptive machine-learning systems, the diffusion of responsibility across developers, vendors, healthcare organisations and clinicians, and the challenge of maintaining effective control where system performance may change following deployment.

The article approaches these questions through a *multi-layered* conception of accountability, drawing principally on the four-part framework developed by Novelli et al. (2024), encompassing *compliance, report, oversight* and *enforcement*. Rather than treating these as discrete or sequential regulatory functions, however, the article examines their *interdependence* across the governance of AI systems. Compliance generates evidence about how AI systems are designed, assessed and operated; reporting makes relevant information available to those responsible for monitoring and decision-making; oversight uses that information to identify risks, allocate responsibility and intervene where necessary; and enforcement provides sanctions or redress while generating information capable of informing subsequent regulatory and organisational responses. The effectiveness of accountability therefore depends not merely on the strength of each layer individually, but on the quality of the connections between them. Accountability gaps are likewise interconnected: weaknesses in one layer can propagate through the governance cycle and undermine the effectiveness of the layers that follow.

The article, after reviewing the whole system and challenges, argues that the AI-related accountability gaps in UK healthcare are structural, distributed and holistic. They are structural because the relevant accountability functions remain institutionally fragmented; distributed because responsibility is shared among stakeholders across the AI supply and deployment chain; and holistic because the effectiveness of one governance layer depends substantially upon the operation of the others. The resulting accountability gaps cannot therefore be addressed adequately by strengthening individual regulatory instruments in isolation. Instead, it requires greater *interoperability* between governance mechanisms so that information and responsibility can flow effectively across the AI lifecycle. The article consequently develops a multi-layered governance perspective in which accountability is understood as a *circular* process rather than a linear sequence of ex ante compliance and ex post enforcement.

The following section, Section 2, develops the conceptual foundation for this analysis by examining the meaning and function of accountability in AI governance. It considers relevant scholarly approaches and competing theoretical perspectives before delving into the four-layer framework of compliance, report, oversight and enforcement drawn from Novelli et al. (2024). Section 3 applies this framework to the UK healthcare system and critically examines the mechanisms through which AI systems are assessed, documented, monitored and controlled. It considers statutory and procurement-based compliance obligations; reporting mechanisms, including safety cases, hazard logs and AIAs; institutional oversight structures involving Clinical Safety Officers (CSOs), NICE and post-market surveillance; and

the legal mechanisms through which sanctions and redress may be pursued. Building on this analysis, Section 4 develops the article's proposed multi-layered governance framework, arguing that closing the accountability gap(s) requires greater interoperability and coordination across the four layers. It reconceptualises accountability as a continuous governance cycle in which evidence generated through one layer informs and strengthens the operation of the others. Section 5 concludes by considering how the UK can recalibrate this governance cycle to support a more integrated and adaptive approach to AI accountability in healthcare, highlighting the need for a strategic leadership.

In sum, this article contributes to the emerging literature on healthcare AI governance in two respects. First, it develops a conceptual account of accountability that moves beyond the identification of discrete regulatory obligations towards an understanding of accountability as an interconnected governance process. Second, it applies this perspective to the UK healthcare context to demonstrate how weaknesses in the connections between governance layers can generate systemic accountability deficits even where individual regulatory mechanisms are formally present. By identifying the *evidence, transparency, attribution* and *learning gaps* that can emerge across these interconnected layers, the article proposes a more integrated and lifecycle-oriented model for the responsible development and deployment of AI in the UK healthcare system.

2 Conceptualising Accountability

2.1 Definitional framework

Accountability is conceptualised as a relational mechanism whereby a decision-maker is required to explain and justify their conduct to an authority or public forum (Bovens, 2007; Novelli et al., 2024; Wieringa, 2020), with the possibility of sanctions for misconduct (Bovens, 2007; Mulgan, 2000). This relationship operates in both internal terms – within a principal-agent dynamic –and external terms, when actors are required to justify their behaviour to those indirectly but significantly affected by their decisions (Bovens, 2007).[4] In this regard, the principal (P) delegates authority to the agent (A) to act on its behalf, with the understanding that the agent will be answerable to a forum (F) for how those interests are represented and pursued. Accountability typically unfolds in three phases: A, after being delegated authority by P, provides information to F; F engages in deliberation and questioning; and, finally, consequences are imposed based on the forum's evaluation (Wieringa, 2020, p. 3).

However, in socio-technical contexts where agency is distributed across human and non-human systems, traditional models are likely to falter. A socio-technical system includes people and organisations in many different roles (for example, developer, manufacturer, user, bystander or policymaker), their interactions, and the processes that organise these interactions (Theodorou and Dignum, 2020). Such a socio-technical account can encompass, amongst others, the algorithmic system's reason for existence, the context of its development, and the effects of the system (Wieringa, 2020, p. 7). As AI systems increasingly mediate high-stakes decisions, such as in healthcare, the need arises for a deeper and more adaptive understanding of accountability that recognises its networked, multi-agent and lifecycle-dependent nature (Wieringa, 2020, p. 10; Novelli et al., 2024, p. 1876).

In contrast to conventional settings, where responsibility may be assigned to a clearly identifiable individual or office (A), algorithmic decision-making involves multiple contributors, including data curators, model designers, software engineers, administrative institutions and end-users. These distributed interactions give rise to what is often called the "problem of many hands", making it difficult to identify a single responsible actor (World Health Organization, 2021, p. 43). Furthermore, systemic features such as bugs, opacity and intellectual property restrictions introduce persistent barriers to accountability, particularly when errors occur or harm is caused (Burrell, 2016; Pasquale, 2016).

2.2 Why AI Accountability Matters: Focus on Healthcare

As AI becomes increasingly embedded in decision-making across sectors, the demand for robust accountability frameworks grows ever more urgent. In many cases and settings, AI operates within socio-technical systems involving algorithms, data infrastructures, institutions and human actors (Ananny and Crawford, 2018; Unver, 2023). These multi-agent systems challenge traditional notions of agency and

[4]Bovens distinguishes between various accountability relationships, including vertical (principal-agent) and horizontal/public (democratic or societal) accountability. He explicitly notes that accountability mechanisms extend beyond formal delegation to include forums indirectly affected by actions (Bovens, 2007, p. 451).

responsibility: outcomes are not solely the product of a single actor's intention but emerge through distributed interactions over time. This requires accountability frameworks that address not only individual liability but also further considerations around 'who-does-what' when incorporating AI.

Shaped by design choices, training data and institutional goals, AI is not neutral and is widely acknowledged to be value-laden (Martin, 2019; Mittelstadt et al., 2016). Given this, AI systems are likely to put fair and equitable access to patient care at risk, absent robust and responsive accountability frameworks. Considering that AI technologies increasingly mediate socially consequential decisions (e.g., from welfare allocation and predictive policing to clinical triage), any attempt to draw an accountability framework should respond to the following questions:

- Who designs, validates and deploys the systems?
- What kinds of data are used and with what limitations?
- Who is accountable when the system causes harm or unjust outcomes? (Martin, 2019)

The above questions do not simply aim to address responsibility and liability but also to resolve ethical concerns from the very early stages, ideally at the design level. Ethical concerns are especially compelling in healthcare, where algorithmic tools are used to diagnose disease, allocate scarce resources, predict patient deterioration and automate administrative decisions. In this context, human lives and well-being are at stake alongside ethical and legal concerns such as bias, discrimination and opacity.

From this vantage point, accountability in AI systems must be a substantive ethical and governance principle, grounded in the rights, needs and dignity of those affected. It cannot be reduced to technical transparency or compliance box-checking (Wieringa, 2020). AI systems thus require new kinds of expertise from legal professionals alongside technical specialists, as these systems need to be able to assess sometimes-conflicting laws and public values, given that they can be vastly complicated and rely on a large amount of connected data sources (Wieringa, 2020).

The importance of accountability is heightened in healthcare, where decisions traditionally rest on strong professional ethics, interpersonal care and a duty of harm avoidance (Naik, Hameed, Shetty et al., 2022). AI systems, unlike human clinicians, are not considered to be moral agents and cannot be sanctioned or held responsible (Unver, 2023). Yet they often operate with a form of functional autonomy, creating what Floridi, Cowls, Beltrametti et al. (2018) call "distributed moral responsibility". This complicates how to address AI accountability in light of both the ethical challenges involved and the need to maintain the principles of biomedical ethics.[5]

2.3 Alternative Models of AI Accountability

Existing literature offers various, often overlapping, models of AI accountability. One group emphasises explainability and transparency, advocating for technical and procedural measures such as Explainable AI (XAI) to render algorithmic systems more interpretable and contestable by affected stakeholders (Doshi-Velez and Kim, 2017; Kaminski, 2023; Nannini, Balayn, and Smith, 2023). Others advocate for ethical design and risk management, drawing on the leading principles of AI ethics and governance, responsible research and innovation (RRI), and value-sensitive design to ensure that ethical considerations are embedded across the AI lifecycle (Dignum, 2019; Hohma, Boch, Trauth, and Lütge, 2023; Umbrello and van de Poel, 2021). A further group promotes governance to ensure actionable accountability, featuring oversight mechanisms, third-party audits and external accreditation processes (Laux, 2024; Mökander and Floridi, 2023; Percy, Dragicevic, Sarkar, and d'Avila Garcez, 2022; Raji, Xu, Honigsberg, and Ho, 2022).

While each approach enriches our understanding of AI accountability, they share a set of recurring limitations:

- *Overreliance on technical tools as proxies for ethical responsibility.* Many models assume that the deployment of XAI, transparency dashboards, or bias audits automatically fulfils moral accountability—ignoring the context in which such tools are used and the broader institutional duties involved (Doshi-Velez and Kim, 2017; Nannini et al., 2023).

[5] T. L. Beauchamp and J. F. Childress (of the Kennedy Centre), writing in Principles of Biomedical Ethics (1979), promulgated four principles intended both to guide and to justify medical decisions. Of these principles, three have ancient origins, beneficence (interventions should always be for the benefit of the patient), non-maleficence (the patient should never be harmed), and justice (which requires fairness and forbids exploitation)). The fourth principle of autonomy, was brought to prominence by the Nuremberg Code, but was expressed long before in the ethics of Immanuel Kant and J. S. Mill and had been given new prominence in the Belmont Report (1978), as requiring respect for persons and their self-determination (Attfield, 2023, p. 83).

- *Narrow legalistic focus insufficiently attuned to socio-ethical concerns.* Some governance approaches treat compliance as the endpoint, centred on meeting legal thresholds such as the AI Act or GDPR, without addressing deeper issues such as moral agency, democratic inclusion or distributional justice (Mökander et al., 2022; Veale and Brass, 2019).
- *Failure to account for institutional complexity and AI lifecycle dynamics.* Accountability frameworks often abstract away from the real-world architecture of organisations that develop, deploy, monitor and wind down AI systems. Lifecycle thinking, covering design, deployment, operation, audit and decommissioning, might be underdeveloped (Batool, Zowghi, and Bano, 2025; Martin, 2019).

Moreover, while the EU has adopted a more centralised, horizontally applicable legislative framework for AI governance, this is not mirrored in the UK, where a principles-based and regulator-led approach has resulted in a more sectoral model of AI governance. In engaging with the ongoing debate concerning AI accountability, it is important to consider not only the resulting fragmentation but also the critical nature of specific domains, such as healthcare, where the stakes are higher and professional cultures of responsibility are already well established. Overall, one could recognise a need for a more integrative framework that reflects the multidimensional nature of accountability, particularly one that is suited to the complexities of the UK healthcare sector, which is the focus of this article.

2.4 Towards a Comprehensive Framework: Novelli, Taddeo & Floridi (2024)

Given the constraints highlighted above, this paper advocates for a holistic governance model that integrates technical tools with meaningful institutional oversight, socio-ethical sensitivity and lifecycle accountability. Novelli et al. (2024) respond to this need by offering a systematic, multi-layered framework that both clarifies the conceptual foundations of accountability and operationalises it for practical governance. Their model defines accountability through three enabling conditions: Authority Recognition, Interrogability and Limitation of Power.

In this context, the Principal (P) grants the Agent (A) authority to serve its interests, while A recognises that the Forum (F) will hold it accountable for how such interests have been served. Limitation of power constrains the arbitrary exercise of delegated tasks (T) by allowing F to monitor A's performance and evaluate its results (Novelli et al., 2024). While providing a broad framework to address the relationships between the involved actors (P, F and A), this model also identifies the key features of the relationship of answerability: context (what for?), range (about what?), agent (who?), forum (to whom?), standards (according to what?), process (how?) and implications (what follows?). A simplified description exemplifies their approach: electoral accountability (context) for choices concerning the recruitment of representatives (range) by political leaders (agent) to the electorate (forum) is assessed against the standard of reliability (standards) through elections (process) and can imply electoral failure (implications) (Novelli et al., 2024). These features orient the practical implementation of accountability mechanisms in different settings, including healthcare.

Furthermore, the framework developed by Novelli et al. (2024) centres on four core accountability goals (or layers):

1. *Compliance*: Ensuring AI systems adhere to existing legal, ethical and professional standards.
2. *Report*: Documenting processes and outcomes in transparent, traceable and accessible ways.
3. *Oversight*: Enabling scrutiny by internal and external bodies across the AI lifecycle.
4. *Enforcement*: Ensuring consequences for breaches of standards, including reparative or punitive measures.

Moving beyond a purely theoretical socio-technical perspective, this framework provides flexibility for both context-specific applications and generalisable dynamics of accountability. It enables examination of the interaction among technology, humans and the wider environment, alongside the capacity to oversee, monitor and evaluate the performance of AI systems embedded and functioning within particular contexts.

Critically, Novelli, Taddeo and Floridi argue that these four goals are not fixed in character but can each be discharged in either a *proactive* or *reactive* mode, with governance actors tending to emphasise one mode over the other depending on their institutional mission. Proactive accountability is implemented ex ante and addresses the question "what should an accountable AI system look like?" by operating through anticipatory standard-setting before a system is deployed or harm occurs (Novelli et al., 2024). Reactive accountability, by contrast, is triggered once a relevant event has occurred; it

concerns "how should the use of an AI system be held accountable?" and is oriented towards addressing the effects of failure after the fact (Novelli et al., 2024). As illustrated in Figure B1 and Figure B2, each accountability mode is represented within this framework as a linear spectrum of sequential steps.

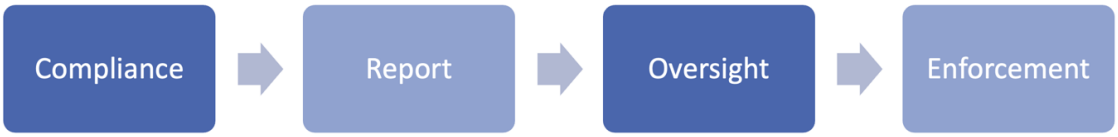

Fig. B1: Proactive Accountability (Adapted from Novelli, Taddeo, and Floridi, 2024)

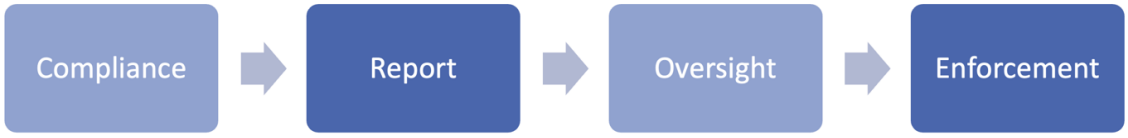

Fig. B2: Reactive Accountability (Adapted from Novelli, Taddeo, and Floridi, 2024)

Novelli et al. (2024) recognise that different AI governance objectives require distinct forms of accountability, arguing that oversight and enforcement are better suited to overarching political missions, whereas compliance and reporting mechanisms are more appropriate for narrower, sector-specific objectives. While this distinction is conceptually valuable, the authors stop short of providing a structured account of how these accountability layers should be understood, integrated and operationalised within a coherent governance model.

The framework's emphasis on both forward-looking (proactive) and backward-looking (reactive) accountability is particularly well suited to clinical environments, where preventing harm is as important as learning from adverse outcomes and evolving socio-technical systems. Nevertheless, the framework presents these dimensions largely as sequential or discrete modes of accountability, offering limited guidance on how they should interact within a comprehensive governance architecture. In response to the need for these mechanisms to function collectively in and across socio-technical contexts, this paper emphasises the *interdependence* of the aforementioned four layers, whereby accountability gaps are minimised through continuous governance cycles and enhanced interoperability.

This paper adopts Novelli, Taddeo and Floridi's framework not merely as a theoretical point of departure, but as the organising axis of its analysis. It argues, however, that the framework requires further development to articulate a more holistic, systemic and integrated conception of AI accountability, as detailed in Section 4.

3 Accountability in Healthcare AI: Critical Analysis of the UK approach

3.1 General Overview of the UK Healthcare System

AI in the UK's healthcare sector is governed through a hybrid regulatory model that combines statutory medical device law with an expanding body of non-statutory standards, professional guidance and institutional practice. This approach reflects the UK's broader 'pro innovation' regulatory philosophy, which emphasises flexibility and sector-specific oversight rather than comprehensive horizontal legislation (Unver and Odusola-Stevenson, 2025). Unlike other sectors, AI policy in healthcare should operate on five cross-sectoral principles, i.e. safety, security and robustness; transparency and explainability; fairness; accountability and governance; contestability and redress, which all regulators must implement within their respective domains and remits (Department for Science, Innovation and Technology (DSIT), 2023). As a consequence, AI systems deployed within clinical environments are primarily governed through existing frameworks – most notably the Medical Devices Regulations 2002, which apply where AI constitutes or functions as a medical device – and through soft law instruments issued by the

NHS England, NHSX (now integrated into NHS England), and professional regulators (MHRA, 2024; NHS England, 2024a).

This decentralised architecture distributes governance responsibilities across four principal bodies, each intervening at a distinct point in operational lifecycle for the AI systems. The MHRA regulates AI as a medical device under the Medical Devices Regulations 2002, requiring UKCA (or, transitionally, CE) marking before market entry and imposing post-market surveillance and vigilance obligations thereafter; its jurisdiction, however, extends only to tools that meet the statutory definition of a 'medical device', leaving many administrative and lower-risk clinical decision-support AI tools outside its regulatory perimeter entirely. NICE, by contrast, does not authorise use but evaluates clinical and cost effectiveness through its Evidence Standards Framework for digital health technologies and, increasingly, its Early Value Assessment pathway, which permits conditional, time-limited NHS adoption pending further real-world evidence; its output is guidance rather than a binding legal gateway, though it carries substantial de facto authority over commissioning decisions. The Care Quality Commission (CQC) operates at a different level again: rather than assessing the AI tool itself, it regulates the safety and quality of the care providers – NHS trusts, general practices, and other registered services – which deploy such tools, under the Health and Social Care Act 2008. Finally, NHS England provides the operational infrastructure through which the other three mechanisms are translated into practice. Its Digital Technology Assessment Criteria (DTAC) act as a gateway for NHS procurement and incorporate two clinical risk management standards, DCB0129 and DCB0160.[6] Notably, the difficulty of navigating this four-body landscape has recently prompted NICE, the MHRA, the CQC and the Health Research Authority to jointly launch the AI and Digital Regulations Service (AIDRS), a coordination portal intended to help developers and NHS adopters identify which obligations apply to a given tool – an institutional response that itself evidences the complexity the present analysis addresses (NHS Health Research Authority, 2023).

While this division of labour across multiple institutions prevents regulatory duplication and supports rapid technological adoption, academic commentary has noted that it also produces a fragmented accountability landscape, particularly with respect to adaptive machine learning (ML) systems whose behaviour evolves post deployment (Cobbe, Singh, and Veale, 2023; Raji et al., 2020). While the UK's principles-based, decentralised model seeks to preserve regulatory flexibility and facilitate innovation, its long-term effectiveness will depend on the enforcement capacities to produce coherent operationalisation of the core principles, specifically safety, security and robustness; appropriate transparency and explainability; fairness; accountability and governance; and contestability and redress. (Unver and Odusola-Stevenson, 2025).

Overall, the UK's approach to the AI governance within healthcare sector is fundamentally decentralised given the sector-led regulatory model and the limited role of the central support functions (DSIT, 2024, paras 9–15; Department for Science, Innovation and Technology (DSIT), 2023, paras 57–62). Arguably, the principal governance challenges include the lack of clear and transparent ownership of governance responsibility within each sector; inconsistency of action and powers between regulatory agencies; and the broader risk that the existing governance ecosystem does not fully and appropriately respond to the risks and challenges posed by AI (McGrath, 2024). As a result, effective healthcare governance in practice depends on the interaction between legal obligations, procurement and assurance processes, clinical safety standards and professional duties of care. Drawing upon the analytical framework articulated by Novelli et al. (2024), this section undertakes a systematic evaluation of the UK healthcare landscape through the lens of compliance, report, oversight and enforcement as the accountability goals.

3.2 Compliance: Regulatory Standards and Procurement Governance

Compliance within the healthcare context necessitates adherence to both medical device regulations and digital safety standards. Under the UK's hybrid model, compliance extends beyond meeting MHRA requirements to include conformity with NHS assurance processes, clinical risk-management obligations,

[6] Although DCB0129 and DCB0160 are formally classified as information standards rather than statutory requirements, they operate as de facto mandatory requirements within NHS digital procurement and deployment: manufacturers are expected to demonstrate compliance with DCB0129, while NHS organisations deploying the technology must undertake the corresponding clinical safety assessment and develop a safety case under DCB0160. NHS organisations cannot lawfully procure a digital health technology without evidence from the manufacturer that it complies with DCB0129, nor can they deploy the technology without a locally developed safety case under DCB0160 (NHS England, 2024b).

data-protection duties and procurement-stage due diligence, collectively shaping a multilayered system intended to safeguard patients while enabling responsible innovation.

3.2.1 Statutory Authority and Classification

In the UK, the MHRA retains statutory primacy over the authorisation and surveillance of medical devices, including those incorporating AI functionalities (MHRA, 2024). Any AI tool performing a medical function – whether diagnostic, prognostic or therapeutic in nature – is classified as a medical device pursuant to the Medical Devices Regulations 2002. The MHRA's updated guidance framework stipulates that manufacturers must furnish a declaration of conformity and maintain rigorous technical documentation as preconditions for registration on the UK market (MHRA, 2022).

3.2.2 Procurement Standards and Institutional Obligations

Healthcare providers (e.g., NHS trusts) bear concurrent obligations to ensure that AI tools satisfy the Digital Technology Assessment Criteria (DTAC) and comply with the DCB0160 standard for clinical risk management prior to deployment (NHS Digital, 2023; NHS England, 2024b). These requirements establish minimum regulatory thresholds for interoperability, cybersecurity and clinical safety. These requirements establish minimum thresholds for interoperability, cybersecurity and clinical safety, though scholarly assessment reveal that their application faces various implementation challenges (Arora and Lawton, 2024; Unver and Odusola-Stevenson, 2025).

3.2.3 The Compliance Gap

Empirical research indicates a growing risk of what may be termed 'procurement-reliance bias', whereby local NHS trusts erroneously assume that central regulatory approval constitutes sufficient evidence of site-specific safety (Vollmer et al., 2020). Although external validation studies may be conducted, a reduction in model accuracy compared with initial development-stage results is generally anticipated (Vollmer et al., 2020). This phenomenon reflects a broader structural weakness in the compliance architecture: the conflation of market authorisation with operational fitness for purpose. This is further complicated by the fact that if ML/AI algorithms do not qualify as medical devices (as not all do), they are not subject to regulatory approval, and thus a large incentive to demonstrate their real world effectiveness is absent (Vollmer et al., 2020). Academic commentary has argued persuasively that compliance must be conceptualised as an 'active' institutional process necessitating continuous engagement, rather than a 'passive' box-ticking exercise confined to vendor selection procedures (Novelli et al., 2024).

3.3 Report: Traceability and Documentation Requirements

Reporting mechanisms facilitate transparency by ensuring that the AI lifecycle is documented for both internal governance review and external regulatory audit, thereby establishing the evidentiary foundation for accountability (Novelli et al., 2024). In the UK healthcare context, such mechanisms include but are not limited to clinical safety case reports, Algorithmic Impact Assessments (AIAs), model cards, ethical/health impact documentation and documentation of human-machine interaction.

Reporting mechanisms facilitate transparency by ensuring that the AI lifecycle is documented for both internal governance review and external regulatory audit, thereby establishing the evidentiary foundation for accountability (Novelli et al., 2024).

3.3.1 Clinical Safety Case Reports

Pursuant to NHS Digital standards, organisations deploying health information technology are required to produce a Clinical Safety Case Report and maintain a Hazard Log throughout the operational lifecycle of the system (NHS Digital, 2023). These documentary instruments constitute the primary reporting mechanism for the systematic identification, assessment and mitigation of algorithmic risks before and after deployment (National Health Service (NHS) Digital, 2021). The Hazard Log, in particular, functions as a living record of identified risks, mitigation strategies and residual uncertainties (NHS Digital, 2023). Nevertheless, concerns persist regarding the adequacy of these conventional documentation mechanisms for capturing the adaptive and emergent characteristics of ML systems, whose behaviour may evolve following deployment.

3.3.2 Algorithmic Impact Assessments

The UK regulatory landscape has progressively moved towards integrating Algorithmic Impact Assessments (AIAs) within healthcare to document potential socio-technical harms, including discriminatory effects across racial, gender and socioeconomic dimensions (Ada Lovelace Institute, 2022). By requiring organisations to identify foreseeable risks before deployment, AIAs strengthen organisational transparency and provide an evidential basis for subsequent regulatory scrutiny. However, the absence of a standardised methodological format frequently results in what critics have characterised as 'opacity by design', whereby proprietary algorithmic architectures remain insulated from meaningful public scrutiny or independent third-party evaluation (Mittelstadt et al., 2016; Pasquale, 2016).

3.3.3 Model Cards and Ethical/Health Impact Documentation

Alongside AIAs, emerging transparency instruments such as Model Cards and Ethical Health Impact Assessments (EHIAs) are increasingly promoted as mechanisms for documenting model design choices, dataset composition, known biases, system limitations and anticipated ethical impacts (Aggarwal et al., 2024). Model Cards provide structured summaries of a model's intended use, performance characteristics and recognised limitations, thereby improving communication between developers, deployers and clinical end users (Gilbert et al., 2025; Mitchell et al., 2019). EHIAs similarly provide a systematic framework for evaluating the broader societal and ethical implications of healthcare AI, including distributive impacts, accessibility concerns and risks of exacerbating existing health inequalities (Ada Lovelace Institute, 2022).

By documenting model behaviour, development provenance and anticipated impacts in a structured and accessible manner, these instruments strengthen traceability and support more informed procurement, deployment and governance decisions (NHS England, 2024a). Although their adoption is not yet mandatory within the UK healthcare system, both academic scholarship and policy guidance increasingly regard them as essential components of a mature AI assurance ecosystem (Gilbert et al., 2025; Government Digital Service, 2025; Lekadir, Frangi, Porras et al., 2025). Their effectiveness, however, ultimately depends upon the extent to which such documentation is standardised, independently verified and incorporated into regulatory and organisational decision-making, rather than treated as a procedural compliance exercise.

3.3.4 Documentation of Human-Machine Interaction

Academic consensus increasingly emphasises that adequate reporting must transcend the purely technical dimensions of the algorithmic 'black box' and comprehensively document the human-machine interaction, with particular attention to the interpretive processes through which clinicians engage with AI outputs and integrate them into clinical decision-making (Azevedo, 2026; Blasimme and Vayena, 2020; Pot and Prainsack, 2024; World Health Organization, 2021). Such documentation enables reviewers to determine whether adverse outcomes arise from algorithmic limitations, inappropriate clinical reliance or wider organisational failures. This requirement reflects recognition that algorithmic accountability cannot be divorced from the sociotechnical context of deployment (Aggarwal et al., 2024; Novelli et al., 2024; Pot and Prainsack, 2024).

3.4 Oversight: Institutional Scrutiny and Monitoring

Oversight constitutes the evaluative dimension of accountability by enabling competent authorities to scrutinise AI-supported decision-making throughout the system lifecycle. Building upon the evidential foundation established through reporting, effective oversight facilitates both *ex ante* assessment of governance arrangements before deployment and *ex post* evaluation of whether AI systems continue to operate consistently with applicable legal, ethical and clinical standards (Novelli et al., 2024).

3.4.1 The Clinical Safety Officer

The primary institutional mechanism for oversight within NHS organisations is the mandatory appointment of a Clinical Safety Officer (CSO), typically a senior clinician possessing both clinical expertise and technical understanding (NHS Digital, 2023). The CSO bears responsibility for assuring the safe deployment and operation of health information technology systems, functioning as the institutional bridge between technical system performance and clinical safety requirements (NHS Digital, 2023). By reviewing safety documentation, monitoring implementation and advising organisational decision-makers, the CSO provides an important mechanism for translating reporting into operational oversight.

Nevertheless, in light of scholarship highlighting the normative, empirical and broader ethical challenges associated with healthcare AI (Morley et al., 2020), questions remain as to whether the CSO role, as presently constituted, possesses sufficient authority, technical expertise and organisational resources to exercise meaningful oversight over increasingly adaptive and complex AI systems.

3.4.2 Post-Market Surveillance: The MHRA Yellow Card System

Oversight within the UK healthcare sector also operates through statutory post-market surveillance mechanisms, most notably the MHRA's Yellow Card Scheme, which provides the national system for reporting and investigating adverse incidents involving medicines, medical devices and AI-enabled software as a medical device (SaMD/AIMD) (MHRA, 2025b). As AI systems increasingly incorporate adaptive or continuously learning functionalities, the Yellow Card Scheme provides an important mechanism for detecting emerging safety risks, performance degradation and previously unidentified harms following deployment.

However, both academic and regulatory analyses identify important limitations. The scheme depends heavily upon voluntary reporting, struggles to identify complex algorithmic failure modes and operates within fragmented AI supply chains in which developers, integrators and healthcare providers each possess only partial knowledge of system behaviour (Vollmer et al., 2020). Consequently, while post-market surveillance provides an indispensable layer of retrospective oversight, it cannot substitute for continuous lifecycle monitoring capable of responding to the dynamic and context-dependent behaviour of AI systems.

3.4.3 External Regulatory Surveillance

External oversight is exercised through several complementary regulatory institutions. CQC evaluates how healthcare providers govern the digital and organisational risks associated with AI-enabled technologies, integrating these considerations into broader assessments of quality and patient safety (Care Quality Commission, 2023). Alongside the CQC, NICE performs an important evaluative function through its assessment of the clinical and cost effectiveness of AI-driven and digital technologies. NICE's Evidence Standards Framework for Digital Health Technologies establishes benchmarks for analytical validity, clinical performance and economic value, and increasingly informs procurement and commissioning decisions throughout the NHS (National Institute for Health and Care Excellence (NICE), 2023). Although NICE does not function as a safety regulator, its recommendations significantly influence the evidential standards that developers must satisfy before technologies are adopted within clinical practice (Wolff and Shaw, 2024).

Complementing these existing mechanisms, the National Commission into the Regulation of AI in Healthcare, established by the MHRA, has recently launched a call for evidence to examine whether existing oversight arrangements remain adequate for adaptive AI systems (MHRA, 2025a). Their findings suggest that the governance of AI-enabled healthcare requires a coordinated, lifecycle-based regulatory approach supported by stronger cross-regulatory collaboration, adaptive post-market surveillance and clearer allocation of regulatory responsibilities (Medicines and Healthcare products Regulatory Agency (MHRA), 2026). Collectively, these institutions illustrate that oversight extends beyond safety monitoring to encompass broader assessments of effectiveness, governance, regulatory coordination and public trust.

3.4.4 The Distributed Oversight Problem

A fundamental challenge for AI governance arises from the distributed nature of contemporary AI ecosystems. AI systems are commonly developed, trained, integrated and deployed through complex algorithmic supply chains involving developers, data curators, infrastructure providers, system integrators, healthcare organisations and clinical end users (Cobbe et al., 2023). This diffusion of responsibility creates what may be termed a 'distributed oversight problem', in which no single entity possesses either the complete information or the requisite authority to exercise comprehensive supervisory control (Novelli et al., 2024).

This article characterises this structural limitation not only unique to oversight layer or mechanisms. Building upon the framework of Novelli et al. (2024), it is argued that effective accountability should be conceived as a coordinated governance function rather than a collection of isolated supervisory activities.

3.5 Enforcement: Liability Frameworks and Mechanisms of Redress

Enforcement represents the remedial dimension of accountability by ensuring that failures to comply with legal, ethical or professional standards result in meaningful consequences. These consequences extend beyond the imposition of civil liability or professional sanctions to include corrective organisational measures, regulatory interventions and mechanisms of redress designed to restore public trust and improve future governance (Novelli et al., 2024).

3.5.1 Legal Liability and Tortious Redress

Within the UK, enforcement remains primarily anchored in traditional tort law and professional regulatory frameworks. Where an AI system contributes to patient harm, liability may be distributed between the manufacturer under product liability principles and the clinician or NHS trust under the law of medical negligence (Price, Gerke, and Cohen, 2024; Smith and Fotheringham, 2020). The Consumer Protection Act 1987 establishes a regime of strict liability for defective products, implementing the former Product Liability Directive 85/374/EEC, although the application of these principles to standalone software and increasingly autonomous AI systems remains legally uncertain and the subject of continuing scholarly debate (Consumer Protection Act 1987, s 1; Twigg-Flesner, 2021). Consequently, existing liability rules continue to rely upon legal concepts that were developed for conventional products rather than adaptive algorithmic systems.

3.5.2 The Enforcement Gap for Adaptive Systems

A significant enforcement gap arises where continuously learning algorithms evolve following deployment, presenting fundamental challenges for fault-based liability frameworks (Froomkin, Kerr, and Pineau, 2019; Marchant and Lindor, 2012; Pasquale, 2022). Existing legal doctrines struggle to attribute responsibility where software behaviour changes over time in response to new data inputs, potentially in ways neither anticipated nor authorised by the original developer (Hartzog, 2015; Ramírez-Barrantes, Gouveia, and Córdova, 2026). This temporal dimension of algorithmic behaviour exposes a structural weakness within conventional enforcement mechanisms, which generally assume that defective behaviour can be attributed to a single identifiable actor or event.

3.5.3 Professional Regulatory Sanctions

Professional regulators, including the General Medical Council and the General Dental Council, consistently emphasise that registered practitioners cannot delegate their professional duty of care to an algorithm, irrespective of the sophistication of the underlying technology (General Medical Council (GMC), 2024, para. 17a; General Dental Council (GDC), 2013, Principle 1). Accordingly, enforcement following AI-related harm frequently focuses upon whether clinicians exercised appropriate professional judgment when relying upon algorithmic outputs (Price, 2018). While this preserves the traditional principle of professional accountability, it also risks obscuring the broader organisational and technological factors that contribute to adverse outcomes, particularly where failures arise from deficiencies in system design, governance or institutional oversight rather than individual clinical decision-making.

4 Closing the Accountability Gap: A Multi-Layered Framework of Governance

The analysis above demonstrates that the UK's governance of AI in healthcare is best understood as a hybrid regulatory ecosystem in which binding statutory obligations operate alongside soft-law instruments, organisational practices and professional norms. While this reflects the UK's pro-innovation regulatory philosophy, it also produces a diffuse distribution of responsibilities that can obscure clear lines of accountability throughout the AI lifecycle. The central question is therefore not whether AI is regulated, but whether the existing governance architecture is capable of filling the accountability gap across increasingly complex socio-technical systems.

This article argues that the accountability gap in UK healthcare AI is *structural*, *distributed* and *holistic*. It is structural because the principal governance functions – compliance, report, oversight and enforcement – are not fully connected institutionally or functionally. It is distributed because responsibility is dispersed across developers, manufacturers, NHS organisations, regulators, healthcare professionals and data controllers, reflecting the "many hands" problem inherent in AI governance (Dixon-Woods and Pronovost, 2016; Froomkin et al., 2019). It is holistic because weaknesses in one layer rarely remain

contained: a defect in the evidence underpinning compliance is likely to propagate, several stages later, into a defect in enforcement outcomes.

A recurring theme emerging from the preceding analysis is that regulatory effectiveness depends less upon the existence of formal legal obligations than upon the institutional capacity of actors responsible for implementing them. Although the MHRA exercises statutory authority over AI-enabled medical devices, operational governance depends equally upon NHS procurement processes, DTAC requirements, clinical safety standards, data protection obligations and the varying levels of digital maturity across NHS trusts. Consequently, the assurance achieved in practice is uneven, increasing the possibility that AI systems may be deployed without sufficiently robust, context-specific validation.

In conjunction with this, the layer of report presents a further systemic vulnerability. Persistent deficiencies relating to algorithmic transparency, model provenance, AIAs, hazard logs and documentation of clinician–AI interactions weaken the evidential chain upon which effective governance depends. Such informational asymmetries impede the ability of regulators and NHS trusts to detect performance drift, reconstruct decision pathways and investigate adverse outcomes, thereby undermining both effective oversight and meaningful enforcement.

These governance challenges become more pronounced for adaptive and continuously learning AI systems. Existing oversight mechanisms, including post-market surveillance through the Yellow Card Scheme, Clinical Safety Officers and local incident reporting, were largely designed for comparatively static technologies and remain only partially calibrated for systems capable of evolving after deployment. As a consequence, oversight becomes distributed across developers, vendors, healthcare providers, clinicians and regulators, none of whom possesses a complete evidential picture of system behaviour throughout its operational lifecycle. The accountability deficit therefore stems not merely from regulatory omission but from the fragmentation of information, responsibility and institutional competence.

Post-market monitoring illustrates this difficulty particularly clearly. Regulatory approval cannot eliminate uncertainty surrounding AI-enabled medical technologies because clinically significant performance drift, reduced reproducibility and changing predictive performance frequently emerge only after deployment in real-world clinical settings. These post-deployment uncertainties expose accountability challenges that extend simultaneously across compliance, report and oversight.

Enforcement remains anchored in traditional tort law, product liability and professional regulation. Although these mechanisms continue to provide important avenues for redress, they encounter increasing difficulty where harm results from adaptive system behaviour, probabilistic outputs or complex interactions between algorithmic recommendations and clinical judgement. Establishing fault, causation and legal responsibility becomes increasingly problematic as AI systems assume greater autonomy, thereby reinforcing the distributed accountability gap identified within the preceding governance layers (Froomkin et al., 2019).

To address these challenges, this article proposes a multi-layered governance framework that conceptualises accountability as an *integrated and continuous governance ecosystem* rather than a collection of discrete regulatory mechanisms. While drawing on the four accountability layers identified by Novelli et al. (2024), the framework advanced here adopts a fundamentally different analytical perspective. Unlike their framework, which conceptualises accountability through distinct reactive and proactive modes, the proposed multi-layered framework reconceptualises accountability as a *circular* governance process. It views compliance, report, oversight, and enforcement not as discrete or sequential stages, but as interdependent components linked through *continuous feedback loops*. Its objective is not merely to classify accountability mechanisms but to close the accountability gap by enabling interoperability between governance layers, ensuring that information generated within one layer continuously informs and strengthens the others. Accountability is therefore understood not simply as an ex ante or ex post legal obligation, but as an ongoing institutional process of coordination, verification and organisational learning.

As illustrated in Figure D3, evidence generated during compliance must feed into reporting mechanisms (Flow 1); reporting outputs must provide meaningful information for organisational and external oversight (Flow 2); oversight findings must support attribution of responsibility and legal redress (Flow 3); and enforcement outcomes must ultimately feed back into regulatory standards, procurement practices and organisational learning (Flow 4). Accountability should therefore be understood as a circular rather than sequential governance process in which each layer depends upon information generated by the preceding layer and simultaneously reinforces those that follow. Only when these reciprocal flows

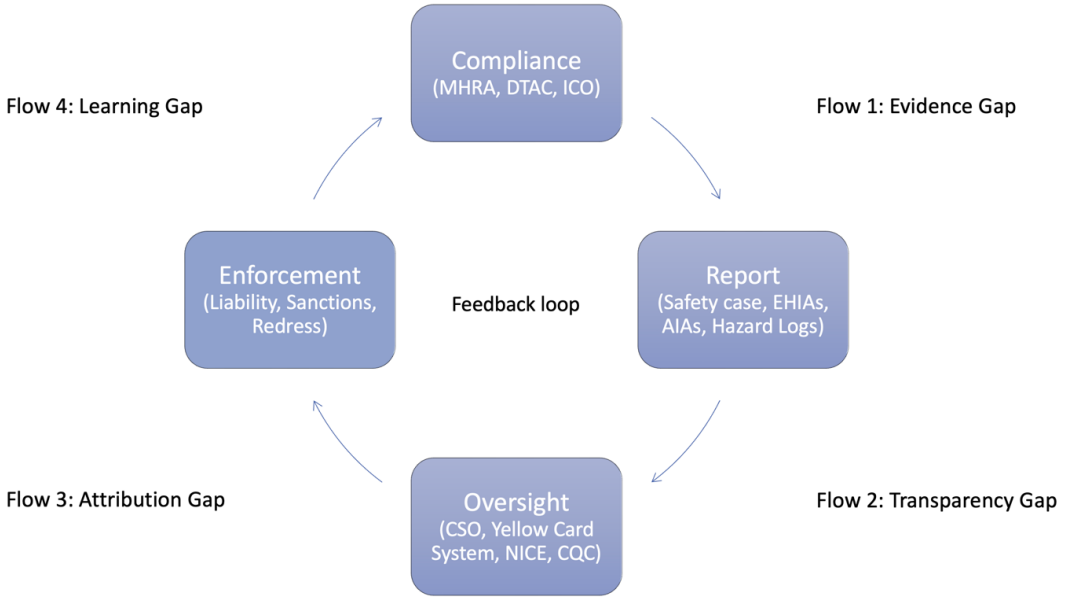

Fig. D3: Feedback loop between interconnected governance layers

operate effectively can the governance system function as a self-correcting, evidence-driven accountability ecosystem. This holistic and iterative architecture is intended to address accountability gaps by strengthening coordination, interoperability, and adaptive governance across the AI lifecycle.

Failures within these flows map directly onto the four categories of accountability gap identified in Table D1. Breakdowns in Flow 1 correspond to the *Evidence Gap*, arising from unrepresentative datasets, hidden assumptions and post-deployment performance drift. Failures in Flow 2 generate the *Transparency Gap*, characterised by inadequate documentation, limited explainability and informational asymmetries. Disruptions in Flow 3 produce the *Attribution Gap*, reflecting the difficulty of assigning legal and organisational responsibility across multiple actors and establishing causation. Finally, deficiencies in Flow 4 result in the *Learning Gap*, where regulatory experiences, adverse events and enforcement outcomes fail to translate into improved standards, procurement practices or institutional learning.

Table D1: Gap Analysis

Evidence Gap	**Transparency Gap**	**Attribution Gap**	**Learning Gap**
Limited representativeness of training and validation data	*Black-box* systems and limited explainability	*Diffused responsibility* across multiple actors	*Weak feedback loops* from incidents to regulation
Performance drift after deployment	Inconsistent reporting and documentation	Difficulty proving causation and fault	Lessons not systematically embedded
Bias and hidden assumptions	*Information asymmetries* between stakeholders	*'Many hands'* problem	Fragmented institutional learning

Read together, Figure D3 and Table D1 demonstrate that the accountability gap is not a single regulatory deficiency but a constellation of interconnected failure modes, each capable of weakening the governance layers that follow. The principal contribution of this framework is therefore to demonstrate that effective accountability cannot be achieved merely by strengthening individual regulatory mechanisms. Rather, it requires governance layers that operate as an accountability ecosystem in which evidence, transparency, attribution and institutional learning continuously reinforce one another. Future governance of AI in UK healthcare should therefore prioritise strengthening these institutional connections and multi-layered interoperability, enabling accountability to function not only as a mechanism

of legal responsibility but also as a continuous process of adaptive governance capable of responding to increasingly autonomous and learning AI systems.

5 Conclusion: Recalibrating the Governance Cycle for the UK

This article has argued that the accountability challenges posed by AI in UK healthcare cannot be understood solely through the adequacy of individual legal or regulatory instruments. Rather, they arise from the interaction between multiple governance layers that remain only partially integrated across the operational lifecycle of AI systems. While the UK's hybrid governance model combines statutory rules with soft-law guidance and organisational governance, its regulatory architecture generates structural vulnerabilities that become particularly significant as AI systems become increasingly adaptive, autonomous and continuously learning.

The analysis demonstrates that accountability in healthcare AI is *structural*, *distributed* and *holistic*. It is structural because compliance, report, oversight and enforcement remain institutionally fragmented; distributed because responsibility is shared across developers, vendors, NHS organisations, regulators and clinicians; and holistic because failures within one governance layer inevitably undermine the effectiveness of those that follow. The accountability gap therefore cannot be addressed by strengthening individual regulatory mechanisms in isolation. Instead, it requires governance capable of sustaining reliable information flows across the entire regulatory ecosystem.

To address this challenge, this article proposes a multi-layered governance framework that reconceptualises accountability as a circular and interoperable process rather than a linear sequence of *ex ante* and *ex post* obligations. While drawing on the four accountability layers identified by Novelli et al. (2024), it adopts a fundamentally different analytical perspective. Rather than distinguishing between proactive and reactive accountability, the proposed framework demonstrates how *compliance, report, oversight,* and *enforcement* function as mutually reinforcing governance layers connected through *continuous feedback loops*. In this respect, whereas Novelli, Taddeo, and Floridi provide a valuable taxonomy of accountability goals or functions, the present framework develops these into an integrated accountability ecosystem centred on their dynamic interaction. It further explains not only where accountability breaks down but also how accountability deficits propagate throughout the governance cycle, conceptualising these failures as the *evidence gap, transparency gap, attribution gap* and *learning gap* across the interconnected governance layers. In doing so, it offers a holistic governance model capable of integrating the legal, technical, and organisational dimensions of AI accountability within a single conceptual architecture.

The analysis further suggests that existing governance arrangements, while necessary, remain insufficient for high-risk AI systems deployed within clinical settings. A more resilient governance ecosystem will require greater standardisation of documentation and transparency requirements, particularly regarding AIAs, safety cases and post-deployment monitoring; enhanced organisational capability through improved AI literacy, procurement expertise and clinical safety culture; regulatory mechanisms capable of addressing continuously learning systems, including update governance, performance drift detection and proportionate re-authorisation pathways; clearer allocation of responsibilities across developers, vendors, NHS providers and clinicians; and liability frameworks better equipped to accommodate system-level harms, emergent model behaviour and complex causal chains (see also Medicines and Healthcare products Regulatory Agency (MHRA), 2026). Collectively, these reforms would strengthen *interoperability* between governance layers and promote accountability throughout the AI lifecycle rather than only after harm has occurred.

Recent institutional developments provide a significant opportunity to advance this agenda. Most notably, the findings of the National Commission into the Regulation of AI in Healthcare reinforce the need for a coordinated, lifecycle-based regulatory framework that extends beyond pre-market approval and embeds continuous oversight throughout the deployment of AI-enabled medical technologies. The Commission emphasises that effective regulation requires adaptive post-market surveillance, robust mechanisms for monitoring model updates and performance drift, greater clarity regarding regulatory responsibilities, and stronger coordination across the UK's healthcare governance landscape (Medicines and Healthcare products Regulatory Agency (MHRA), 2026). These recommendations closely align with the accountability architecture developed throughout this article and provide important institutional support for a more integrated approach to AI governance.

Nevertheless, institutional innovation alone will not resolve the structural fragmentation that characterises the current governance landscape. The Commission itself recognises that the multiplicity of

regulators, guidance frameworks and assurance mechanisms creates complexity for both developers and healthcare providers, underscoring the importance of governance arrangements capable of coordinating regulatory functions across the AI lifecycle (Medicines and Healthcare products Regulatory Agency (MHRA), 2026). Effective accountability therefore depends not merely on the existence of regulatory institutions, but on *strategic coordination* that enables compliance, report, oversight and enforcement to operate as mutually reinforcing governance layers rather than isolated regulatory activities.

In this respect, a stronger leadership role for the Secretary of State for Health and Social Care would help provide the strategic coherence necessary to support an *interoperable accountability ecosystem*. Consistent with the Secretary of State's overarching statutory responsibility to promote a comprehensive health service under the National Health Service Act 2006, central government is well placed to convene relevant regulators, promote common information standards, facilitate cross-regulatory information sharing and strengthen institutional coordination across the AI lifecycle. The recent findings of the National Commission into the Regulation of AI in Healthcare based on its Call for Evidence (Medicines and Healthcare products Regulatory Agency (MHRA), 2026) also reveal the significance of such a strategic coordination. From this standpoint, we believe, the effective use of these existing statutory powers would facilitate the implementation of the proposed multi-layered accountability framework, enabling a more coherent governance for the AI systems.

Overall, as AI systems continue to evolve, accountability must likewise evolve beyond fragmented obligations towards a dynamic governance cycle in which evidence generated through *compliance* informs transparency through *report*, report strengthens attribution through *oversight*, and oversight enables learning through effective *enforcement*. Only through such an integrated governance model can the UK sustain public trust, safeguard patient safety and realise the transformative potential of AI in healthcare.

Declarations

All authors certify that they have no affiliations with or involvement in any organisation or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Ada Lovelace Institute (2022) Algorithmic impact assessment: A case study in healthcare. Ada Lovelace Institute, URL https://www.adalovelaceinstitute.org/project/algorithmic-impact-assessment-healthcare/, accessed: 2026-08-08

Aggarwal N, et al (2024) Artificial intelligence in healthcare. In: Bullock JB, et al (eds) The Oxford Handbook of AI Governance. Oxford University Press, Oxford, p 838–859

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. 10.1177/1461444816676645

Arora A, Lawton T (2024) Artificial intelligence in the NHS: Moving from ideation to implementation. Future Healthcare Journal 11(3):100183

Attfield R (2023) Applied Ethics: An Introduction. Polity, Cambridge

Azevedo M (2026) What would you prefer: A real doctor or a chatbot? a thought experiment on authority, responsibility, and AI in medicine. In: Gouveia S (ed) The Palgrave Handbook on the Ethics of Artificial Intelligence. Palgrave Macmillan, Cham, p 501–514, 10.1007/978-3-032-15112-4

Batool A, Zowghi D, Bano M (2025) AI governance: A systematic literature review. AI Ethics 5:3265–3279. 10.1007/s43681-024-00653-w

Blasimme A, Vayena E (2020) The ethics of AI in biomedical research, patient care, and public health. In: Dubber MD, Pasquale F, Das S (eds) The Oxford Handbook of Ethics of AI. Oxford University Press, New York, p 703–718

Bovens M (2007) Analysing and assessing public accountability: A conceptual framework. European Law Journal 13(4):447–468. 10.1111/j.1468-0386.2007.00378.x

Burrell J (2016) How the machine “thinks”: Understanding opacity in machine learning algorithms. Big Data & Society 3(1):1–12. 10.1177/2053951715622512

Care Quality Commission (2023) Using technology to support care. Guidance

Cobbe J, Singh J, Veale M (2023) Algorithmic harms and digital markets: The socio-technical chains of machine learning. arXiv preprint, URL https://arxiv.org/abs/2304.14749, accessed: 2026-06-01

Department for Science, Innovation and Technology (DSIT) (2023) A pro-innovation approach to AI regulation. CP 815. UK AI White Paper 2023, URL https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper, accessed: 2026-08-08

Dignum V (2019) Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Springer, Cham, 10.1007/978-3-030-30371-6

Dixon-Woods M, Pronovost P (2016) Patient safety and the problem of many hands. BMJ Quality & Safety 25(7):485–488. 10.1136/bmjqs-2016-005232

Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608, accessed: 2026-08-08

DSIT (2024) Government response to the AI white paper consultation. DSIT, URL https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response, accessed: 2026-08-08

Floridi L, Cowls J, Beltrametti M, et al (2018) AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines 28(4):689–707. 10.1007/s11023-018-9482-5

Froomkin AM, Kerr I, Pineau J (2019) When AIs outperform doctors: Confronting the challenges of a tort-induced over-reliance on machine learning. Arizona Law Review 61:33–99

General Dental Council (GDC) (2013) Standards for the dental team. GDC, URL https://www.gdc-uk.org/standards-guidance/standards-and-guidance/standards-for-the-dental-team, accessed: 2026-08-08

General Medical Council (GMC) (2024) Good medical practice. GMC, URL https://www.gmc-uk.org/professional-standards/good-medical-practice-2024, accessed: 2026-08-08

Gilbert S, et al (2025) Could transparent model cards with layered accessible information drive trust and safety in health AI? npj Digital Medicine 8(1):1–5. 10.1038/s41746-025-01482-9

Government Digital Service (2025) Algorithmic transparency recording standard: Guidance for public sector bodies. URL https://www.gov.uk/government/publications/guidance-for-organisations-using-the-algorithmic-transparency-recording-standard/algorithmic-transparency-recording-standard-guidance-for-public-sector-bodies, accessed: 2026-08-08

Hartzog W (2015) Unfair and deceptive robots. Maryland Law Review 74:785–826

Hohma E, Boch A, Trauth R, et al (2023) Investigating accountability for artificial intelligence through risk governance: A workshop-based exploratory study. Frontiers in Psychology 14:1073686. 10.3389/fpsyg.2023.1073686

Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nature Machine Intelligence 1(9):389–399

Kaminski ME (2023) Understanding transparency in algorithmic accountability. In: Barfield W (ed) The Cambridge Handbook of the Law of Algorithms. Cambridge University Press, Cambridge, p 121–140

Laux J (2024) Institutionalised distrust and human oversight of artificial intelligence: Towards a democratic design of AI governance under the european union AI act. AI & Society 39:2853–2866. 10.1007/s00146-023-01777-z

Lekadir K, Frangi AF, Porras AR, et al (2025) FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ 388:e081554. 10.1136/bmj-2024-081554

Marchant GE, Lindor RA (2012) The coming collision between autonomous vehicles and the liability system. Santa Clara Law Review 52:1321–1340

Martin K (2019) Ethical implications and accountability of algorithms. Journal of Business Ethics 160:835–850

McGrath C (2024) The state and values of AI governance in UK healthcare. In: Handbook of Health, AI and the Law. URL https://www.ncbi.nlm.nih.gov/books/NBK613210/, accessed: 2026-08-08

Medicines and Healthcare products Regulatory Agency (MHRA) (2026) National commission into the regulation of AI in healthcare: Research, engagement and call for evidence findings. URL https://www.gov.uk/government/publications/national-commission-into-the-regulation-of-ai-in-healthcare-research-engagement-and-call-for-evidence-findings, accessed: 2026-08-08

MHRA (2022) Implementation of medical devices future regime. URL https://www.gov.uk/government/publications/implementation-of-the-future-regulation-of-medical-devices, accessed: 2026-08-08

MHRA (2024) Software and AI as a medical device. URL https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device, accessed: 2026-08-08

MHRA (2025a) Call for evidence: National commission into the regulation of AI in healthcare. URL https://assets.publishing.service.gov.uk/media/69d3747a019a4faf2745b354/Regulation_of_AI_in_Healthcare_-_information_and_questions.pdf, accessed: 2026-06-01

MHRA (2025b) Medical devices: Post-market surveillance. URL https://www.gov.uk/government/collections/medical-devices-guidance-for-manufacturers-on-vigilance, accessed: 2026-08-08

Mitchell M, et al (2019) Model cards for model reporting. In: Proceedings of FAT*, pp 220–229, 10.1145/3287560.3287596

Mittelstadt BD, et al (2016) The ethics of algorithms: Mapping the debate. Big Data & Society 3(2):1–21. 10.1177/2053951716679679

Mökander J, Floridi L (2023) Operationalising AI governance through ethics-based auditing: An industry case study. AI and Ethics 3:451–468. 10.1007/s43681-022-00171-7

Mökander J, et al (2022) Conformity assessments and post-market monitoring: A guide to the role of auditing in the proposed european AI regulation. Minds and Machines 32(2):241–268. 10.1007/s11023-021-09577-4

Morley J, et al (2020) The ethics of AI in health care: A mapping review. Social Science & Medicine 260:112649. 10.1016/j.socscimed.2020.113172

Mulgan R (2000) "accountability": An ever-expanding concept? Public Administration 78(3):555–573

Naik N, Hameed B, Shetty D, et al (2022) Legal and ethical consideration in artificial intelligence in healthcare: Who takes responsibility? Frontiers in Surgery 9:862322

Nannini L, Balayn A, Smith AL (2023) Explainability in AI policies: A critical review of communications, reports, regulations, and standards in the EU, US, and UK. arXiv preprint arXiv:2304.11218, URL https://arxiv.org/abs/2304.11218, accessed: 2026-08-08

National Health Service (NHS) Digital (2021) DCB0129 clinical risk management: Its application in the manufacture of health IT systems. URL https://digital.nhs.uk/data-and-information/information-standards/governance/latest-activity/standards-and-collections/dcb0129-clinical-risk-management-its-application-in-the-manufacture-of-health-it-systems, accessed: 2026-08-08

National Institute for Health and Care Excellence (NICE) (2023) Evidence standards framework for digital health technologies. URL https://www.nice.org.uk/what-nice-does/digital-health/evidence-standards-framework-esf-for-digital-health-technologies, accessed: 2026-08-08

Nemitz P (2018) Constitutional democracy and technology in the age of artificial intelligence. Philosophical Transactions of the Royal Society A 376:20180089. 10.1098/rsta.2018.0089

NHS Digital (2023) DCB0160 clinical risk management: Its application in the deployment and use of health IT systems. URL https://digital.nhs.uk/data-and-information/information-standards/governance/latest-activity/standards-and-collections/dcb0160-clinical-risk-management-its-application-in-the-deployment-and-use-of-health-it-systems/, accessed: 2026-08-08

NHS England (2024a) Artificial intelligence guidance for IG professionals. URL https://digital.nhs.uk/data-and-information/information-governance/guidance/artificial-intelligence-guidance-for-ig-professionals, accessed: 2026-06-01

NHS England (2024b) Digital technology assessment criteria (DTAC): Guidance for buyers and suppliers. URL https://transform.england.nhs.uk/key-tools-and-info/digital-technology-assessment-criteria-dtac/, accessed: 2026-08-08

NHS Health Research Authority (2023) One-stop-shop for artificial intelligence (AI) and digital regulations for health and social care launched. URL https://www.hra.nhs.uk/about-us/news-updates/one-stop-shop-artificial-intelligence-and-digital-regulations-health-and-social-care-launched/, accessed: 2026-08-08

Novelli C, Taddeo M, Floridi L (2024) Accountability in the governance of artificial intelligence: A four-part framework. AI & Society 39:1871–1882

Pasquale F (2016) The Black Box Society. Harvard University Press, Cambridge, MA

Pasquale F (2022) Liability standards for medical robotics and AI. In: DiMatteo LA, Poncibò C, Cannarsa M (eds) The Cambridge Handbook of Artificial Intelligence: Global Perspectives on Law and Ethics. Cambridge University Press, Cambridge, p 200–212

Percy C, Dragicevic S, Sarkar S, et al (2022) Accountability in AI: From principles to industry-specific accreditation. AI Communications 34:181–196. 10.3233/AIC-210080

Pot M, Prainsack B (2024) Governing AI technologies in healthcare: Beyond the "ethics bubble". In: Handbook on Public Policy and Artificial Intelligence. Edward Elgar Publishing, p 411–422, 10

.4337/9781803922171.00039

Price IW. Nicholson (2018) Medical malpractice and black-box medicine. In: Cohen IG, Fernandez Lynch H, Vayena E, et al (eds) Big Data, Health Law, and Bioethics. Cambridge University Press, Cambridge, p 295–306

Price IW. Nicholson, Gerke S, Cohen IG (2024) Liability for use of artificial intelligence in medicine. In: Solaiman B, Cohen IG (eds) Research Handbook on Health, AI and the Law. Edward Elgar, Cheltenham, p 150–166

Raji ID, Xu P, Honigsberg C, et al (2022) Outsider oversight: Designing a third-party audit ecosystem for AI governance. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22), pp 557–571

Raji ID, et al (2020) Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. arXiv preprint arXiv:2001.00973, accessed: 2026-08-08

Ramírez-Barrantes R, Gouveia S, Córdova C (2026) The role of AI in biomedicine: Computer vision and digital pathology in cancer diagnosis. In: Gouveia S (ed) The Palgrave Handbook on the Ethics of Artificial Intelligence. Palgrave Macmillan, Cham, 10.1007/978-3-032-15112-4_30

Smith H, Fotheringham K (2020) Artificial intelligence in clinical decision-making: Rethinking liability. Medical Law International 20(2):131–154

The Alan Turing Institute (2023) Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. URL https://www.turing.ac.uk/news/publications/understanding-artificial-intelligence-ethics-and-safety, accessed: 2026-08-08

Theodorou A, Dignum V (2020) Towards ethical and socio-legal governance in AI. Nature Machine Intelligence 2(1):10–12

Twigg-Flesner C (2021) Guiding principles for the updating the product liability directive for the digital age. URL https://europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ELI_Guiding_Principles_for_Updating_the_PLD_for_the_Digital_Age.pdf, accessed: 2026-08-08

Umbrello S, van de Poel I (2021) Mapping value sensitive design onto AI for social good principles. AI Ethics 1(3):283–296

Unver MB (2023) Governing fiduciary relationships or building up a governance model for trust in AI? review of healthcare as a socio-technical system. International Review of Law, Computers & Technology 37(2):198–226

Unver MB, Odusola-Stevenson S (2025) UK AI governance at a crossroads: Charting the path ahead. International Journal of Mass Communication 3(2):168–183

Veale M, Brass I (2019) Administration by algorithm? public management meets public sector machine learning. International Data Privacy Law 9(3):190–200

Vollmer S, et al (2020) AI in healthcare: Critical questions. BMJ 368:l6927

Wachter S, Mittelstadt B, Floridi L (2017) Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law 7(2):76–99

Wieringa M (2020) What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*), 10.1145/3351095.3372833, accessed: 2026-08-08

Wolff S, Shaw J (2024) Digital health technology evaluation: Accountability, evidence and public value. Health Policy and Technology 32:101123

World Health Organization (2021) Ethics and governance of artificial intelligence for health: WHO guidance. URL https://www.who.int/publications/i/item/9789240029200, accessed: 2026-08-08

Zerilli J, et al (2019) Transparency in algorithmic and human decision-making: Is there a double standard? Philosophy & Technology 31(4):661–683