FRONTIER ARTIFICIAL INTELLIGENCE AND HEALTH TECHNOLOGY ASSESSMENT

REPORT BY THE DECISION SUPPORT UNIT

February 2026

Authors: Harry Hill[1], Emily Pulsford[1], Allan Wailoo[1], Nicholas Latimer[1]

[1]School of Health and Related Research, University of Sheffield

Decision Support Unit, SCHARR, University of Sheffield, Regent Court, 30 Regent
Street, Sheffield, S1 4DA

Tel (+44) (0)114 222 0734
E-mail dsuadmin@sheffield.ac.uk
Website www.nicedsu.org.uk
Twitter @NICE_DSU

ABOUT THE DECISION SUPPORT UNIT

The Decision Support Unit (DSU) External Assessment Group is based at the University of Sheffield with members at York, Bristol, Exeter, Leicester, Warwick, Swansea and the London School of Hygiene and Tropical Medicine. The DSU is commissioned by The National Institute for Health and Care Excellence (NICE) to provide a research and training resource to support the Institute's Centre for Health Technology Evaluation Programmes. Please see our website for further information www.nicedsu.org.uk.

The production of this document was funded by the National Institute for Health and Care Excellence (NICE) through its Decision Support Unit. The views, and any errors or omissions, expressed in this document are of the authors only. NICE may take account of part or all of this document if it considers it appropriate, but it is not bound to do so.

This report should be referenced as follows:

Haill H, Pulsford E, Wailoo A, Latimer N. Frontier artificial intelligence and health technology assessment. NICE DSU Report, 2026.

EXECUTIVE SUMMARY

This report examines how NICE's health technology assessment (HTA) methods, processes and evidence requirements may need to evolve to assess frontier artificial intelligence (AI) used in health and care. It draws on a range of literature: horizon scanning of AI health technology documents, evaluation frameworks, health economic methods, adoption and implementation evidence, and regulatory-adjacent guidance to identify where existing HTA approaches are under strain, and to propose practical, proportionate adaptations that can be embedded within NICE's current processes. Throughout, we focus on distinguishing where frontier AI creates genuinely different HTA challenges compared with other technologies, and where it may amplify challenges NICE already faces.

What is meant by Frontier AI

Frontier AI refers to AI technologies whose effects, risks and costs depend strongly on how they are configured, integrated and governed in practice, and which change in meaningful ways after deployment. The defining feature is not novelty or size, but whether the technology behaves as a configurable and evolving service rather than a fixed intervention, such that performance cannot be separated from implementation choices and lifecycle management.

A concrete example used in the report is an AI 'copilot' embedded in electronic patient record systems to support clinical documentation, information retrieval or workflow coordination. Although often presented as a single product, its real-world impact depends on how it is set up: what data it can access, how outputs are framed, whether clinicians must review and approve outputs, and whether the system can trigger downstream actions. Updates may be delivered remotely and frequently, meaning that the same named system can behave differently over time. This illustrates why frontier AI can require a more explicit lifecycle approach in HTA, with clearer versioning and review triggers, because the intervention may change over time.

Why frontier AI creates challenges for HTA

Many of the issues raised by frontier AI have close parallels in NICE's existing work. Uncertainty is universal in HTA; heterogeneity, changing comparators, evolving clinical practice and implementation variability are familiar; and NICE already appraises technologies where value depends on pathway context, service organisation and real-world delivery. In that sense, frontier AI does not introduce an entirely new set of methodological categories.

However, in our view, frontier AI goes beyond these familiar challenges in two important ways. First, it makes post-deployment change not just more likely, but an inherent and often rapid feature of the technology: while comparators, populations and practice evolve over time for any intervention, the intervention itself may evolve continuously through routine updates to models and prompts, changes in data sources and safeguards, and modifications to workflow integration and operating configuration. This raises the probability that effectiveness, safety, equity and cost-effectiveness alter after guidance is issued, even when the technology appears to be the same product. Second, these changes can be distributed, incremental and less visible, occurring across multiple system components and sometimes without a clearly labelled new version, making it harder to detect when evidence is no longer aligned with real-world use. As a result, the core HTA challenge is not simply greater uncertainty at appraisal, but a higher risk of misalignment between what was evaluated and what is deployed, unless evidence maintenance, change governance and review triggers are built in.

Key challenges

A central challenge is intervention instability. For frontier AI, change is routine rather than exceptional. Systems may evolve through updates to models, prompts, data sources, safeguards or workflow integration, raising basic questions about what exactly has been appraised and whether the evidence still applies once the system is in use. This leads to evidence decay. Even high-quality evidence may lose relevance as systems change, making review processes based solely on fixed time intervals poorly aligned with risk. For frontier AI, the question is often not whether evidence was once valid, but whether it remains valid.

The report identifies weak or inconsistent system specification and transparency as a problem. High-level product descriptions often fail to capture configuration choices, workflow roles and dependencies that directly shape safety, effectiveness, equity and cost of Frontier AI. Without clearer specification, HTA would struggle to establish what is being assessed, to compare appraisals, or to ensure that recommendations apply to real-world use.

Comparator choice is particularly difficult because frontier AI often reshapes workflows or redistributes tasks rather than substituting cleanly for existing practice. Comparator selection is therefore a substantive design choice, not a technical detail, and strongly conditions estimates of value and cost-effectiveness.

Transferability and generalisability cannot be assumed. Frontier AI performance depends on local data quality, digital infrastructure, staffing, training and governance. Treating transferability as implicit risks overstating value or underestimating risk when technologies are deployed across diverse National Health Service (NHS) settings.

The report also highlights sociotechnical risk and human–AI interaction as central. Benefits and harms depend on how clinicians interpret outputs, how uncertainty is communicated, and how oversight are designed. Technical accuracy alone is an unreliable guide to real-world impact.

Frontier AI can also act as a disruptive technology, altering care pathways, workforce roles and service models. As a result, HTA must consider system-wide effects and knock-on consequences, not just local performance improvements.

From a health economic perspective, frontier AI introduces dynamic and structural uncertainty. Pricing and use may change over time; monitoring, governance and assurance generate ongoing costs; health system productivity gains may not translate into savings; and induced demand can alter costs across pathways. Static, point-in-time economic models therefore risk misrepresenting both value and affordability.

Finally, the report identifies additional cross-cutting issues, such as organisational readiness, procurement and cybersecurity burden, accountability across complex value chains, data governance across jurisdictions, and the reliability of the evidence trail for multi-use systems that affect whether HTA decisions on frontier AI can be implemented consistently and remain valid over time.

Recommendations

The recommendations in this report respond to these challenges in a practical way that NICE can apply in guidance and appraisal. They are designed to strengthen the ability to evaluate and compare the technology, ensure recommendations remain aligned with real-world use as systems evolve.

1. **Specify precisely what the intervention is that is to be evaluated.** Frontier AI cannot be reliably assessed using high-level product descriptions. A structured specification defines the AI system-in-use, its components, configuration, workflow role and boundaries, so that appraisal is anchored to what is actually deployed and evidence can be interpreted consistently.
2. **Bind NICE guidance to a specific version, configuration and scope of use, with predefined review triggers.** This addresses evidence decay by making clear when changes are significant enough to require review, rather than assuming recommendations apply indefinitely to a technology that continues to evolve.
3. **Require Predetermined Change Control Plans or equivalent HTA change plans.** Frontier AI systems are likely to change after deployment through updates to the model/version, prompts, data sources, integrations, workflow configuration, or safety safeguards. Therefore, appraisal should include how such changes will be managed, validated and communicated. Making this explicit reduces the risk that untracked updates change the intervention in use, weakening confidence that the original evidence and assurance still apply.
4. **Strengthen transparency requirements.** Because frontier AI is often marketed or perceived as broadly capable, there is a risk it is used beyond the specific configuration and use case that has been evaluated. NICE may therefore require a standardised set of disclosures that clearly sets out the evaluated system’s intended purpose and pathway role, appropriate use and misuse risks (how it should and should not be used); known limitations and failure modes; subgroup performance and key evidence gaps; and how uncertainty is communicated to users. This would help keep implementation

within the limits of the assessed evidence, supporting safe use and consistent deployment across settings.

5. **Add a short frontier AI checklist to NICE guidance to ensure consistent reporting.** This is not a recommendation for NICE to assess cybersecurity or legal compliance. Instead, a frontier-AI annex would require sponsors/manufacturers to set out, clearly and consistently, the regulatory and assurance arrangements already in place (including cybersecurity and wider governance), alongside other implementation-related safety risks. This would allow NICE to judge whether the evidence is transferable and the technology can be used safely in practice, while relying on existing AI governance and cybersecurity oversight rather than duplicating it.

6. **Treat the human–AI system as the intervention, rather than the AI component alone.** Safety and effectiveness depend on interaction design, training, oversight and behavioural responses, particularly for systems that support reasoning, documentation or workflow coordination.

7. **Strengthen NICE's capacity to assess frontier AI and standardise multidisciplinary assessment processes.** Frontier AI raises clinical, technical, organisational and behavioural questions, so consistent access to multidisciplinary expertise is useful to avoid variability in decisions and delays in appraisal.

8. **Require structured transferability analysis and UK-relevant real-world evidence planning.** For frontier AI, performance and risk are highly sensitive to local context, so appraisal must include how evidence will translate into NHS settings and how performance will be monitored over time.

9. **Strengthen comparator expectations and enforce explicit comparator justification at scoping.** Frontier AI often reconfigures pathways rather than replacing existing practice in a simple way, making comparator selection central to credible value estimation.

10. **Take account of the real work and costs of putting frontier AI into practice.** Even if a product meets the right cybersecurity and governance requirements, there can still be significant effort and expense to buy it, connect it to NHS systems, keep it secure, and maintain assurance over time—and this can vary a lot between sites. NICE should reflect these

practical burdens in costs and feasibility, because they can determine whether something that looks cost-effective on paper can actually be rolled out and sustained safely at scale.

11. **Strengthen economic evidence to reflect lifecycle use and uncertainty for high-update frontier AI.** This is not a call to model expected price changes over time. Rather, economic analyses should reflect how costs and effects may change with updates, changing utilisation, and the ongoing requirements for monitoring, governance and assurance. This would help ensure conclusions come from plausible real-world deployment patterns, rather than relying on a single, static snapshot of costs and performance.
12. **Use proportionate conditional recommendation pathways and maintained guidance for high-update frontier AI.** Linking access to evidence generation, monitoring and explicit stop criteria allows NICE to support timely innovation while retaining oversight as systems evolve.

Taken together, these recommendations aim to ensure that NICE can continue to support valuable AI innovation while maintaining safety, equity and value for money, by anchoring appraisal to real systems in use, recognising lifecycle change as intrinsic, and strengthening how implementation and governance considerations are incorporated within HTA for frontier AI, complementing NICE's existing surveillance and post-guidance functions.

Relative feasibility and practicality of the recommendations for NICE implementation

Although the recommendations are intended to operate as a coherent package, they differ in how readily they can be implemented within NICE's current HTA processes and institutional constraints. In the authors' judgement, some recommendations largely formalise or extend approaches NICE already uses in other contexts (and could therefore be adopted relatively quickly through updated guidance, templates and committee practice). Others would require new forms of evidence, clearer conventions for defining and tracking system change over time, stronger links with developers and deployers, and more active post-recommendation oversight (which may not be solely within NICE's remit). For clarity, we group them into three broad tiers of relative feasibility.

Tier 1 (high feasibility, near-term): The most readily implementable recommendations are those that strengthen established HTA practices rather than introducing new evaluative domains. This includes strengthening comparator expectations by enforcing explicit comparator justification at scoping (Recommendation 9); strengthening economic evidence to reflect lifecycle use and uncertainty for high-update frontier AI (Recommendation 11); treating the human–AI system as the intervention rather than the AI component alone (Recommendation 6); and strengthening relevant disclosures so committees and implementers are clear on the evaluated use case, limitations and uncertainty (Recommendation 4).

Tier 2 (moderate feasibility, targeted development): A second group of recommendations fits NICE's direction of travel, but would benefit from piloting and clearer shared conventions. For frontier AI, recommendations often can't be treated as a one-off judgement, because the same AI tool can behave differently as versions change, configurations differ, and it is embedded into local workflows. We therefore recommend making appraisal and guidance more tightly tied to the real system being used in practice by binding guidance to version/configuration/scope and setting clear review triggers (Recommendation 2); requiring structured transferability analysis and a UK-relevant real-world evidence plan to confirm (or revise) applicability in NHS use (Recommendation 8); using proportionate conditional recommendations and maintained guidance where updates are frequent (Recommendation 12); and adding a short frontier AI checklist/addendum to standardise minimum reporting across NICE routes (Recommendation 5). These build on existing lifecycle and evidence-generation approaches, but are more operationally demanding because they require consistent definitions of the AI system-in-use, workable conditions for reappraisal, and practical monitoring and reassessment arrangements that do not create disproportionate burden to NICE appraisal teams.

Tier 3 (lower feasibility, longer-term): This includes requiring structured frontier AI system specifications as a condition of appraisal (Recommendation 1); requiring Predetermined Change Control Plans or equivalent HTA change plans and mechanisms to track and assess updates between formal reviews (Recommendation 3); and taking account of the real work and costs of procurement, integration, assurance and ongoing maintenance in economic and feasibility judgements (Recommendation 10). Strengthening NICE's internal capacity to assess frontier AI and standardising multidisciplinary assessment processes is also best viewed as a longer-term aspiration (Recommendation 7): it is central to consistency across appraisals, but resource-intensive and dependent on institutional development.

CONTENTS

TABLES

ABBREVIATIONS AND DEFINITIONS

ABHI	Association of British HealthTech Industries
AI	Artificial Intelligence
AXREM	Association of Healthcare Technology Providers for Imaging, Radiotherapy and Care
DSIT	Department for Science, Innovation and Technology
DSU	Decision Support Unit
EHR	Electronic Health Record
FDA	Food and Drug Administration (United States)
HTA	Health Technology Assessment
LLM	Large Language Model +1
LMM	Large Multi-modal Model
MAIC	Matching-Adjusted Indirect Comparison
MHRA	Medicines and Healthcare products Regulatory Agency (United Kingdom)
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NIHR	National Institute for Health and Care Research
NLP	Natural Language Processing
OECD	Organisation for Economic Co-operation and Development
PCCP	Predetermined Change Control Plan
STC	Simulated Treatment Comparison
TSD	Technical Support Document
WHO	World Health Organization

1. Introduction

This document has been prepared for NICE to support its work on how health technology assessment (HTA) methods, processes and evidence expectations may need to adapt for frontier AI used in health and care. It meets the tender requirements by: (i) explaining how frontier AI differs from earlier digital health technologies and more narrowly defined AI applications; (ii) setting out the main technical, evidentiary, procedural, ethical, organisational and economic challenges that frontier AI creates for HTA; and (iii) proposing practical and proportionate actions that NICE could adopt, test and refine over time.

The search strategy for this targeted review is described in the Appendix. In reviewing the literature, we examined how authors define AI and frontier AI, the evaluation challenges they identify for AI technologies, and our assessment of whether and how those challenges are more pronounced for frontier AI. This included considering features such as rapid updating, sensitivity to context of use and performance that depends on how systems are deployed in practice. We also identified practical adaptations to HTA methods and processes that could help NICE respond to these challenges without requiring wholesale changes to existing frameworks.

Findings at the level of individual studies are presented in summary tables in the Appendix. In the main text, we synthesise the evidence to highlight recurring issues that are most relevant for NICE decision-making. To guide the reader, we first provide a brief overview of the types of studies and sources included. We then set out a working definition of frontier AI, followed by the key challenges it poses for HTA, and conclude with recommendations for how HTA approaches could evolve in response.

2. Summary of retrieved studies

A targeted evidence identification strategy was undertaken to support the NICE Frontier AI targeted review. The objective of this strategy was to identify relevant empirical evidence, guidance, and policy documents addressing the evaluation, regulation, and real-world use of frontier and adaptive artificial intelligence systems in healthcare.

The approach combined structured bibliographic database searching with targeted grey literature identification. The appendix provides full details of the database search terms and strategy. Bibliographic searches were conducted in MEDLINE, Embase, and EconLit using validated artificial intelligence search filters, supplemented with additional terms designed to capture frontier and adaptive AI concepts, including generative models, autonomous systems, continuous learning, model updating, and post-deployment change. Searches were not restricted by publication date, geography, or study design. Title and abstract screening was undertaken to identify literature relevant to frontier or adaptive AI applications in health and care settings.
In parallel, targeted grey literature searching was undertaken to identify regulatory guidance, policy documents, methodological frameworks, and horizon scanning outputs relevant to AI-enabled medical technologies. This included targeted searching of websites of regulatory authorities, health technology assessment agencies, international organisations, professional bodies, and policy institutions with established activity in AI governance and evaluation. Supplementary searching was undertaken to identify key reports and guidance frequently cited in this domain but not indexed in bibliographic databases.

In total, 37 documents were reviewed in full and included in the final reference list generated through this process. These documents comprise 17 guidance, policy, regulatory, or framework publications, including outputs from national and international regulators, health technology assessment bodies, professional associations, and policy organisations; 18 peer-reviewed empirical studies or reviews reporting on AI system performance, evaluation approaches, economic considerations, or methodological issues; and 2 preprints reporting early empirical evidence relating to highly autonomous or agentic AI systems and clearly identified as non-peer-reviewed. This distribution reflects the current maturity of the evidence base for frontier and adaptive AI in healthcare, in which governance, regulatory, and methodological development currently outpaces the availability of formal evaluative and outcomes-focused research on interventions specifically termed Frontier AI. Guidance documents originate from the United Kingdom, United States, Canada, international organisations, and European institutions. These guidance documents primarily address artificial intelligence technologies exhibiting frontier-like characteristics, such as adaptivity, continuous learning, autonomy, or post-deployment model change. In contrast, empirical studies and reviews are fewer in number and are more internationally distributed.

This report draws on a wide range of evidence relevant to assessing frontier AI for health technology assessment. This includes horizon scanning, evaluation frameworks, adoption and implementation perspectives, methodological research, and regulatory and policy guidance. Key sources include NIHR Innovation Observatory horizon scanning on AI and generative AI technologies (Oyewole et al., 2021; Lanyi et al., 2025), NICE's Evidence Standards Framework for Digital Health Technologies and its expectations around performance over time and real-world evidence (NICE, 2022), and the LSE evaluation framework for professional-facing digital health and AI technologies (van Kessel et al., 2025). We also draw on evidence about adoption and delivery challenges from industry bodies such as the Association of British HealthTech Industries (ABHI, 2023; ABHI, 2025), and on imaging, cybersecurity and procurement perspectives from the Association of Healthcare Technology Providers for Imaging, Radiotherapy and Care (AXREM, 2025a; AXREM, 2025b).

Methodological work on adapting HTA approaches for AI-based medical devices and clinician-facing tools provides further foundations, including studies addressing assessment domains, evidence requirements and value frameworks for AI technologies (Farah et al., 2024; Boverhof et al., 2024; Di Bidino et al., 2024; Jacob et al., 2025). Regulatory-adjacent guidance on lifecycle governance and transparency for machine learning–enabled medical devices also informs the analysis, particularly with respect to managing change over time and maintaining assurance after deployment (FDA, 2025; MHRA/FDA/Health Canada, 2024). In addition, broader discussions of frontier AI governance highlight the importance of post-deployment evaluation, transparency and monitoring in settings where systems evolve rapidly and are difficult to observe directly (DSIT, 2025; Bommasani et al., 2025). The purpose of bringing these sources together is to translate a diverse evidence base into a coherent set of concepts and actions that are practical, proportionate and workable within existing NICE processes, while identifying where those processes may need to adapt for frontier AI.

This report also draws on international guidance and horizon scanning that specifically address large multi-modal models and generative AI in health. WHO's ethics and governance guidance highlights the inherent unpredictability of these systems over their lifecycle, and emphasises the responsibilities of developers, providers and deployers, the need for workforce training and public engagement, post-release auditing, and clear approaches to liability (World Health Organization, 2024). These issues are directly relevant to HTA because they influence real-world effectiveness and safety, as well as the ongoing costs and feasibility of governance arrangements for frontier AI.

Finally, the analysis incorporates cross-country policy insights that are particularly relevant to general-purpose AI systems, where rapid evolution and wide applicability increase the risk of fragmented, informal or inequitable adoption if governance does not keep pace (OECD, 2024). A 2025 horizon scan watch list from Canada's Drug Agency identifies AI technologies likely to diffuse in the near term, including AI tools for clinical documentation, training and education, disease detection and diagnosis, treatment support and remote monitoring, alongside cross-cutting issues such as privacy, liability and accountability, bias and data quality, data sovereignty and environmental impact (Canada's Drug Agency, 2025). Together, these sources help identify the issues most likely to be relevant when NICE considers the value, affordability and acceptable risk of frontier AI for the NHS.

Detailed information on the search strategy, search terms, and included study summaries is provided in the appendix. The appendix summary tables present structured information on how frontier or adaptive features are defined or implied, the key evaluation and governance challenges identified, and corresponding solutions either reported by study authors or developed by the review team. The appendix also includes illustrative case examples drawn directly from the included studies to demonstrate how these challenges may arise in real-world healthcare settings.

3. A definition of frontier AI for HTA purposes

For HTA purposes, we recommend frontier AI to be defined as AI technologies whose effects on care, costs and risk cannot be separated from how they are configured, integrated and governed in practice, and which can change in meaningful ways after deployment. It follows that a technology should be treated as frontier AI when it behaves less like a fixed product and more like an evolving service, such that evidence, performance and value depend on ongoing implementation choices and lifecycle management.

In health and care, frontier AI typically refers to systems built on large, general-purpose AI models that have been trained on very large and diverse datasets. Rather than being designed for a single, narrow task from the outset, these models are adapted for different uses through configuration choices such as how they are instructed, what information they are allowed to draw on, or how they are connected to other software systems, instead of being rebuilt as bespoke tools for each task (Moor et al., 2023).
For NICE HTA, the key question is therefore not the size of the model or the sophistication of the technology, but whether the AI is likely to have system-level effects. A technology should be considered frontier AI when it has the potential to influence multiple steps of a care pathway, change how work is shared between staff, and shape outcomes and resource use in ways that depend strongly on how the system is set up, governed and maintained over time. In these cases, performance cannot be separated from implementation, configuration and ongoing oversight. What matters for HTA is whether the combination of capabilities, ongoing change and governance needs makes the technology’s effects unstable, context-dependent, and difficult to evidence or reproduce without clearly defining how it is actually used in

practice. This challenges standard HTA assumptions about fixed interventions, stable effects and evidence that remains valid over time. Where it is unclear whether such system-level effects will arise, NICE can require targeted evidence on how the system changes over time, how it is monitored, and what risks are managed, to determine whether it should be treated as fixed AI or frontier AI (Moor et al., 2023).

Several characteristics tend to underpin these system-level effects. Frontier AI often produces a range of outputs, such as summaries, draft notes or suggested next steps, rather than a single, fixed result. Outputs may vary even when the same information is provided, which makes testing, validation and monitoring more challenging. Performance can also change depending on how users interact with the system, what information it is allowed to access, and how it is built into clinical workflows. In addition, these systems are often updated frequently through cloud-based services or software connections, sometimes more often than traditional medical devices are updated, which increases the importance of managing change over the lifecycle of the technology (FDA, 2025; MHRA/FDA/Health Canada, 2024). Frontier AI can also introduce specific risks, such as generating incorrect but convincing information, encouraging over-reliance by users, mishandling sensitive data, or allowing errors to spread across connected workflows, all of which require active monitoring after deployment (DSIT, 2025; Bommasani et al., 2025). WHO guidance on large multi-modal models similarly highlights unpredictable outputs, new uses emerging over time, and changing performance as features that are important for governance (World Health Organization, 2024).

A practical example that fits this definition is an electronic patient record–integrated clinical copilot built on a general-purpose AI model, such as systems currently being piloted and deployed by major health record vendors and health systems. For example, Epic Systems has integrated large language model based tools into its electronic health record to support clinical documentation, summarise patient histories and assist with workflows during clinical encounters, including drafting notes and preparing follow-up actions as part of routine care (Moor et al., 2023). These systems do not operate as standalone decision tools. Instead, they combine a general-purpose model with access to patient records, local clinical guidance and software functions that allow outputs to be written into the health record or trigger actions within existing workflows.

In practice, how such a system behaves depends on many configuration choices: which data it can access, how outputs are guided, what safety controls are in place, and how it fits into everyday clinical work. These elements are often updated remotely through cloud services or software interfaces, meaning that the same deployed system can change in reliability, behaviour or influence on care over time, even if it is described as the same product. For this reason, regulatory and policy analyses increasingly describe frontier AI as a continuously configured service rather than a static medical device, with performance and risk shaped by the interaction between the technology and its context of use (FDA, 2025; MHRA, FDA, & Health Canada, 2024).

This is why, for frontier AI, HTA cannot focus only on evaluating a fixed algorithm at a single point in development. Instead, appraisal needs to consider how the system is used in practice and how it is allowed to change once deployed, including how updates are managed, checked and communicated, and who is responsible if performance or downstream clinical effects change over time (Moor et al., 2023; FDA, 2025; MHRA, FDA, & Health Canada, 2024).

By contrast, a technology that would not usually be considered frontier AI is a locked, task-specific model with a single, clearly defined output, such as a locally installed chest X-ray triage tool that flags pneumothorax suspected (yes/no), with stable versioning and infrequent, controlled updates. Such tools typically affect a single step in care rather than reshaping whole pathways, and their evidence base is more likely to remain valid across settings and over time under standard HTA assumptions (Moor et al., 2023; Elvidge et al., 2024).

4. Frontier map: what is likely to come next

NIHR horizon scanning identifies a pipeline of generative AI–enabled technologies in clinical development that are beginning to move beyond narrowly bounded analytic tasks, particularly in areas where text generation and interaction with clinical information are central. The most mature applications identified are LLM-based systems that generate clinical text, including tools for drafting clinical notes and discharge summaries, generating recommended treatment plans or dose suggestions, and providing conversational support to patients through virtual assistants, especially in mental health and cancer-related pathways. These technologies are predominantly classified as digital tools rather than devices, are often disease-agnostic, and are among the few generative AI applications reported at higher technology readiness levels, suggesting they may be among the earliest to reach routine use. Other applications in the pipeline include image-generation systems intended to enhance or annotate diagnostic images, data generation and augmentation tools designed to enrich training datasets or simulate clinical data, and 3D model generation technologies that reconstruct anatomical structures from imaging to support procedural planning. While most technologies remain at early or intermediate stages of development and are unevenly specified in terms of model architecture, the scan indicates that their potential effects are likely to arise not from isolated outputs, but from how these systems are incorporated into documentation workflows, clinical decision-making processes, and patient interaction over time (Lanyi et al., 2025).

The foundation-model literature, exemplified by Moor et al. (2023), anticipates progress toward generalist medical AI systems built on large, reusable foundation models rather than task-specific algorithms. These systems are characterised by three core technical properties: the ability to accept and combine multiple medical data modalities (such as imaging, clinical text, laboratory results, physiological signals, and structured EHR data), the capacity for dynamic task specification through natural-language prompts rather than retraining, and the use of explicit medical knowledge representations to support reasoning and explanation. On this basis, the paper outlines a set of concrete application areas that such systems could support, including drafting grounded radiology reports that link textual findings to image regions, interactive clinical documentation tools that generate notes from conversations and EHR context, bedside decision-support systems that summarise patient state and provide explanatory alerts or recommendations, and augmented procedural support that integrates visual, textual, and anatomical information during surgical or endoscopic workflows. Importantly, these examples are presented as anticipated applications enabled by model capabilities, rather than as fully deployed technologies, and the authors emphasise that their flexibility, adaptability, and capacity to take on previously unseen tasks challenge existing assumptions about validation, regulation, and evaluation of medical AI as a single, stable function (Moor et al., 2023).

The Canada Watch List offers a practical near-term indication of where frontier-like AI is most likely to emerge first within health care settings, by identifying five technology areas expected to have a significant impact over the next five years: AI for notetaking; AI tools to accelerate and optimize clinical training and education; AI for disease detection and diagnosis; AI for disease treatment; and AI for remote monitoring. Across these areas, the report repeatedly emphasises impacts on administrative burden, workflow efficiency, clinician workload and burnout, and the organisation of care, rather than only improvements in a single clinical endpoint.

For example, under AI for notetaking, the Watch List describes AI-powered notetaking applications that use automatic speech recognition and natural language processing to transcribe clinician–patient conversations and generate clinical notes that clinicians then review and sign. It gives concrete examples including AI scribes (producing transcripts, medical notes, and referral letters) and reports quantified workflow effects from evaluations (e.g., reductions in administrative time, reduced after-hours work, and improved job satisfaction), while also flagging risks such as errors/omissions and hallucinations that require clinician review. It also names other tools such as PhenoPad (an open-source note-taking interface capturing free-form notes and standardised phenotypic data via speech, NLP, and handwriting recognition) and Tali, which integrates AI scribes with dictation and medical information retrieval to streamline documentation.

Under clinical training and education, the report describes tools that can summarise evidence and support upskilling/reskilling, and provides examples including

OpenEvidence (a medical language model that aggregates/synthesises clinically relevant evidence and provides citations), ChatGPT used for tasks such as simulated patient scenarios, quizzes, critique of communications, and summarising research, and AI-VSP, which combines AI with virtual reality to create immersive simulated patients and personalised training experiences.

For disease detection and diagnosis, the Watch List notes that many AI/ML-enabled devices to date focus on detection/diagnosis and highlights radiology and other specialties, giving examples such as ASIST-TBI, designed to screen head-injury CT scans in emergency settings to rapidly identify traumatic brain injury and support earlier neurosurgical escalation, and LumeNeuro, which uses polarimetric retinal imaging and machine learning to detect biomarkers associated with neurodegenerative disease. It also explicitly raises system effects such as the risk that diagnostic AI could increase downstream demand for tests and follow-up interventions, potentially straining capacity if not targeted appropriately. For disease treatment, the Watch List defines roles including identifying optimal treatment plans (e.g., medication choice/dose, considering interactions, tailoring to patient profiles) and assisting triage/risk assessment, and gives examples spanning different modes of treatment AI: Kaia Health (a digital therapeutics company using machine learning to deliver interventions for conditions including musculoskeletal pain, COPD, and osteoarthritis), Wysa (an AI mental health chatbot supporting CBT programmes and on-demand therapist support), and Valence Labs' LOWE, described as an LLM-orchestrated workflow engine for executing complex drug-discovery workflows in natural language.

Finally, for remote monitoring, the Watch List describes systems combining biomedical sensors (e.g., wearables/smartphones/smart home sensors) with AI/ML analytics to generate alerts, predict adverse events, and support proactive intervention. It provides examples including AlayaCare (remote patient monitoring plus documentation and portals; a cited Canadian clinical study reports reductions in emergency visits and hospitalisations in COPD/CHF cohorts during a three-month implementation) and Coughy (AI-based sound analysis using digital audio biomarkers from smartphone/smartwatch cough recordings for real-time remote cough monitoring). The report also notes that the net impact depends heavily on measurement accuracy and highlights that false positives/negatives or overdiagnosis could increase burden and cost.

DSIT's discussion paper highlights a further change likely to define the next phase of frontier AI: the integration of models with tools and scaffolds that enable multi-step, goal-directed behaviour, including planning, task sequencing, and execution with reduced human prompting (DSIT, 2025). The paper does not point to specific frontier AI healthcare technology case studies, but its implications for health care are clear. Rather than systems that simply generate discrete outputs, this trajectory suggests AI that coordinates actions across records, communication channels, and decision workflows. The DSIT emphasises that these emerging systems exhibit brittleness alongside apparent competence. That is, Frontier AI may produce plausible but incorrect outputs, fail to follow instructions reliably, or behave unpredictably when tasks involve multi-step reasoning, tool use or interaction with other digital systems (DSIT, 2025). As these characteristics enter clinical settings, performance observed in controlled evaluations may not fully anticipate behaviour under real-world complexity, time pressure and incomplete information.

A credible near-term evolution is the movement from single-model 'copilots' toward modular, role-specialised multi-agent architectures that coordinate multi-step clinical tasks. In urgent-care settings, Hayat et al. (2025) evaluate a multi-agent, large language model–based clinical system designed to autonomously carry out substantial parts of a virtual care encounter. The system consists of multiple role-specialised agents that collectively conduct patient history-taking through dialogue, synthesise information across the interaction, generate structured clinical documentation, produce a differential diagnosis, and propose investigations and treatment plans prior to clinician review. In contrast to a single copilot model that responds to prompts, this architecture coordinates a sequence of interdependent clinical tasks with limited human prompting, effectively shaping the flow and content of care delivery. Studied across 500 real-world virtual urgent-care encounters, this approach has potential effects on workforce allocation, consultation throughput, and clinical standardisation, while also concentrating risk in how errors, omissions, or biases introduced early in the interaction may propagate through subsequent steps of the care pathway.

In procedural and surgical domains, Oettl et al. (2025) describe AI systems that are increasingly embedded within integrated surgical ecosystems, combining software with robotic platforms, imaging, navigation, and intraoperative monitoring technologies. These systems use AI to support preoperative planning, guide instrument positioning and navigation during procedures, adapt actions based on real-time data, and monitor surgical progress. Rather than operating as stand-alone decision-support tools, such technologies function as coordinating layers across hardware, software, and clinical workflows. Their deployment may improve procedural precision, consistency, and efficiency, but also introduces new dependencies between algorithmic behaviour, device performance, and human oversight, complicating accountability, safety assurance, and evaluation of clinical and economic value.

WHO's guidance on large multimodal models anticipates continued expansion of LMM-based applications across diagnosis and clinical care, patient-facing functions, clerical work, education and research. It also cautions that rapid adoption and persuasive, human-like interaction may outpace institutional preparedness, creating risks to trust, rights and safe use if governance and training do not evolve in parallel (World Health Organization, 2024). This suggests a near-term pipeline of LMM-based technologies that, while not frontier AI systems in themselves, can generate frontier-AI-like effects through their scale, adaptability, and integration across clinical, administrative, and patient-facing functions, with impacts distributed across clinical quality, workforce experience, patient interaction, and system trust rather than concentrated in a single outcome domain.

Regulatory and governance developments indicate that planned change and lifecycle transparency are becoming baseline assumptions for frontier systems. FDA's PCCP framework and the trilateral transparency principles reflect an expectation that AI systems will evolve after deployment, and that such evolution must be visible, structured, and subject to oversight (FDA, 2025; MHRA/FDA/Health Canada, 2024). These developments are relevant because they reflect the technological reality of frontier AI systems as a class; namely, that post-deployment change is anticipated, even though the regulatory documents themselves do not point to specific technology case studies. Similarly, policy discussions of frontier AI consistently frame the core challenge as operating under persistent uncertainty and partial opacity, with transparency, third-party evaluation, and post-deployment monitoring positioned as mechanisms for governing systems whose behaviour cannot be fully characterised in advance (Bommasani et al., 2025; DSIT, 2025). Finally, frontier AI is increasingly expected to act as a disruptive technology, reshaping delivery models, skill mix and organisational boundaries across multiple pathways rather than incrementally improving a single step. INAHTA emphasises that such technologies tend to blur the line between intervention and system redesign, with impacts that unfold over time and across organisational levels (INAHTA, 2022). OECD analysis reinforces that if these dynamics are not anticipated, fragmented and inequitable adoption patterns may become entrenched, increasing cost and undermining trust (OECD, 2024).

Taken together, these trajectories suggest that the near-term frontier is defined less by discrete, named technologies and more by configurable, evolving systems embedded within clinical workflows such as multi-agent virtual care platforms, AI-supported surgical and navigation systems, and LMM-based clinical coordination tools, whose behaviour, impact, and risk profile emerge through use rather than being fixed at launch. In a context where the literature indicates that real-world deployments remain limited, understanding what frontier AI is likely to become, rather than focusing narrowly on the few early implementations, is likely a preferable approach to inform the design of HTA approaches that remain relevant as these technologies enter routine health-system practice.

5. Key challenges for HTA posed by frontier AI

Drawing on our targeted review and synthesis, the challenges below reflect our judgement about what is most likely to matter for NICE in practice, rather than a simple restatement of issues listed in the literature. NICE’s established appraisal approaches rest on the assumption that the technology under assessment can be specified with sufficient clarity and stability for evidence generated during development and evaluation to be interpreted as evidence about what will be deployed, used, and reimbursed in practice. Frontier AI directly challenges this assumption. Because frontier AI systems are general-purpose, configurable, and capable of evolving through updates to models, data sources, prompts, guardrails, and workflow integrations, the object of appraisal is not a fixed intervention but a system whose behaviour is contingent on configuration and context of use.

The resulting uncertainty is therefore not simply greater in magnitude, but qualitatively different in kind. It concerns what the technology is at any given point in time, what functions and actions it is permitted to perform within clinical workflows, how it is actually used in practice, whether evidence remains valid following routine updates or reconfiguration, and whether observed outcomes can be attributed to the AI system itself rather than to the surrounding sociotechnical system in which it is embedded. This shared problem of defining, evidencing and attributing value to a moving and context-dependent intervention provides the common foundation for the specific HTA challenges set out below.

6. Intervention instability and the problem of defining the object of the evaluation

Frontier AI poses a fundamental challenge to a core assumption of NICE health technology assessment: that the intervention under appraisal can be described in a stable enough way for evidence generated during development and evaluation to remain a reliable guide to what will be used in practice. For frontier AI, change is not an occasional exception but an expected feature of how these systems operate. Unlike most pharmaceuticals, devices, or earlier digital health technologies, frontier AI can change through several connected parts of the overall service, rather than through a single, clearly labelled product update. Changes to any of these parts can plausibly affect effectiveness, safety, equity and resource use after appraisal (FDA, 2025; MHRA, FDA, & Health Canada, 2024).

In practice, the deployed system may change in multiple ways: the underlying AI model may be updated; the information it draws on may be altered (for example, which documents or data sources it uses); the instructions and templates that shape its outputs may be revised; the safety controls that constrain outputs or determine escalation may be modified; and the way it is presented to users or fitted into clinical workflows may be adjusted. Taken together, these routes mean that what is described as the same product can behave differently over time, even when there is no obvious change in the stated model version. This raises a foundational HTA question: what, precisely, is the intervention that NICE has appraised and recommended?

A comparable issue is already familiar in HTA for complex service-level interventions or care pathway redesigns, where outcomes depend not only on the intervention concept but on how it is implemented, integrated and delivered in practice. In those settings, small changes in delivery or context can alter effectiveness and resource use, even when the headline intervention appears unchanged. Frontier AI raises the same underlying problem, but more sharply, because changes can occur more frequently and are often less visible, making it harder to be confident that evidence remains aligned with what is used in practice.

An illustrative example of this challenge is the application of generalist medical AI. These can perform a wide range of clinical tasks across different data types, depending on how they are prompted, what information they retrieve, and how they are configured. Moor et al. (2023) describe how these models enable broad medical functionality, while also challenging traditional approaches to validation and regulation, because task boundaries are fluid and system behaviour is not fixed in advance. From an HTA perspective, this creates blurred boundaries around intended use and makes it difficult to anchor evaluation to a single, stable function or outcome. Appraisal therefore cannot rely on broad claims about capability. Instead, it requires clearly defined descriptions of how the system is actually used, explicit roles within clinical workflows, and safeguards to prevent the system being used in ways that have not been evaluated. Ongoing monitoring and clear rules about updates are also needed, because evidence may no longer apply as the system changes over time.

7. Evidence decay and the durability of HTA conclusions

A direct consequence of instability in frontier AI systems is that evidence can become out of date as a normal part of use, rather than only in rare or exceptional circumstances. Even high-quality evidence may lose relevance if the system that is actually deployed changes in ways that affect outcomes, costs or who benefits. WHO guidance on large multi-modal models highlights that risks may emerge during real-world use because outputs can be unpredictable, new uses may arise over time, and performance may change, even when development and provision are well managed (World Health Organization, 2024).

For NICE, this means that appraisal and review processes may need to assume that change is likely, rather than treating it as an exception. Decisions about when to review guidance and update evidence could therefore be linked to the risk that the system's behaviour or performance has changed. This includes the possibility of gradual or unintended changes arising from software updates, changes in the data the system relies on, or evolving patterns of use, rather than relying solely on fixed, calendar-based review schedules.

The FDA's guidance on Predetermined Change Control Plans provides a practical example of how this approach can work in practice. It requires developers to set out in advance which types of changes are expected, how those changes will be checked and tested, what standards must still be met, and how any effects on safety and performance will be assessed, with the aim of maintaining confidence over time (FDA, 2025). The trilateral transparency principles make a similar point, emphasising the need for ongoing communication so that users and other stakeholders understand what has changed, what remains within scope, and how performance and limitations are monitored throughout deployment (MHRA/FDA/Health Canada, 2024).

WHO guidance on large multi-modal models emphasising that at scale, errors and harm are inevitable and must be anticipated through post-release auditing, published impact assessments, and clear arrangements for liability and redress (World Health Organization, 2024). This strengthens the case for NICE to treat lifecycle evaluation and monitoring as standard requirements for high-impact frontier AI, rather than as optional extras.

A comparable issue already arises in HTA more generally because evidence is always generated in a particular time and context. Even for conventional technologies evaluated in randomised trials, the relevance of results can erode over time as comparators change, treatment pathways evolve, populations treated change, and outcome measurement or clinical practice develops. Real-world and observational evidence raises similar durability issues: conclusions remain valid only as long as the underlying data continue to reflect current practice, and changes in treatment patterns, coding practices or patient populations can reduce the relevance of observational estimates over time. Frontier AI sits within this broader HTA reality, but adds a distinctive additional source of evidence decay: in addition to changing comparators and contexts, the intervention itself is more likely to change through routine updates and reconfiguration. This means that, for frontier AI, the risk of misalignment between the evaluated system and the deployed system is higher, and maintaining relevance requires active review not only of evolving clinical context but also of changes to the system-in-use itself.

8. Version ambiguity and defining significant change

A further and closely related challenge is determining what should count as a significant or substantial change in a way that is workable and proportionate for HTA. This challenge is particularly acute for frontier AI because material changes in system behaviour can arise through multiple update routes such as changes to the model, data sources, prompts, guardrails or workflow integration, without a clearly identifiable new version in the traditional medical device sense. As a result, it can be unclear when the technology NICE has appraised has changed sufficiently that the evidence base, recommendations or reimbursement conditions no longer apply to what is being used in practice.

Expert discussions of adaptive AI regulation highlight persistent uncertainty about where change thresholds should be set and the implications of those choices. Thresholds determine accountability (who is responsible for detecting, declaring and initiating re-review), the feasibility and burden of reassessment, and the incentives created by governance rules. Thresholds that are too rigid risk encouraging informal or silent updates designed to avoid triggering reassessment, while overly permissive thresholds weaken assurance, transparency and trust by allowing potentially value-altering changes to proceed without scrutiny (Aquino et al., 2024). In frontier AI, where change is frequent and often incremental, these tensions arise routinely rather than in exceptional cases.

A related difficulty is that, even beyond frontier AI, NICE has historically found it operationally challenging to determine when guidance should be revisited in practice. Technology appraisal guidance previously often stated that recommendations would be reviewed after a specified period, but in practice such reviews were relatively rarely undertaken and could be difficult to prioritise against other demands. NICE's developing whole-lifecycle direction of travel and more explicitly proportionate approach to evidence and review aim to address this, but the underlying problem remains: it is not straightforward to know when new information is sufficiently important to justify reassessment. This is particularly true when emerging evidence is expected to come from real-world registries or routine data sources that are not consistently analysed over time. In such cases, it may be unclear whether there is new evidence until substantial analytic effort is invested, creating a practical detection problem for review triggers as well as a resource burden.

For NICE, this supports treating significant change not as a purely technical or regulatory label, but as a relevant concept linked to recommendation conditions. It would not be realistic to pre-define every change that might matter for frontier AI, especially when updates are frequent and impacts may only become clear over time. A workable approach is therefore to pre-specify a short list of changes that should always prompt review (for example, changes to intended use/claims, target population or setting, the tool's role in the pathway or degree of autonomy, major integration changes with other tools or electronic health records functions, or changes to safety mitigations), and to complement this with a clear rule that review should also be triggered by what is seen in real-world use (for example, new safety signals, subgroup performance concerns, or clear changes in resource use). This follows the same basic HTA logic (revisit decisions when something changes that could alter value, risk or feasibility) while recognizing the pace and uncertainty of frontier AI.

9. Continuous configuration, post deployment drift and lifecycle accountability

These challenges are intensified by the fact that many frontier AI systems function as ongoing services rather than fixed products, with meaningful changes occurring through routine adjustments rather than clearly labelled software updates. Changes may be made to the information the system draws on, the way outputs are guided or constrained, or how the system is embedded in clinical workflows, and these changes may be introduced by vendors or NHS organisations without being presented as a new version in the traditional sense. As a result, the behaviour of the deployed system can change over time even when the nominal model version appears unchanged, meaning that evidence may lose relevance because what is being used in practice is no longer fully aligned with what was originally assessed.

A comparable dynamic is already familiar to NICE technology appraisals where recommendations are made for a specific place in the treatment pathway, but the pathway itself evolves over time as new appraisals recommend technologies at earlier, later, or adjacent lines of therapy. As the sequence of available options changes, the original standard care comparator and the incremental value of the originally appraised technology can change, potentially making earlier cost-effectiveness conclusions less applicable even if the original technology has not changed. Frontier AI raises an analogous problem, but through a different mechanism: instead of the pathway changing around a stable product, the product and the surrounding sociotechnical system can change within the same nominal deployment, with implications for both effectiveness and cost-effectiveness. In both cases, the underlying challenge is that NICE guidance is anchored to a decision problem that may not remain stable, and conclusions can become misaligned if the relevant pathway context changes.

NIHR horizon scanning suggests that publicly available information about how AI systems are configured and operate is often limited, reducing transparency and increasing the risk that meaningful functional changes occur without being visible to HTA bodies, commissioners or users (Forsythe et al., 2023; Forsythe et al., 2024; Lanyi et al., 2025). As a result, proposals for governing generative AI place increasing emphasis on layered forms of assurance, including clear records of changes, independent scrutiny, and well-defined responsibility for monitoring performance and learning from incidents after deployment, recognising that users may not be able to detect when a system’s behaviour has changed in ways that matter (Azad, 2025).

These concerns are even more pronounced for continuous-learning or unlocked AI-enabled medical devices, which are explicitly designed to adapt based on data generated during routine use. CADTH notes that in such cases one-off pre-market evidence is often insufficient, because performance may change over time and may not transfer consistently across settings or populations. Ongoing monitoring and predefined limits on acceptable performance therefore become central to maintaining reasonable assurance of safety and effectiveness over the lifecycle of the technology (CADTH, 2022).

For NICE, this implies that some frontier AI appraisals may need to assess not only baseline performance but also the update mechanism itself as part of the intervention, including what aspects of the system may change, under what conditions, with what oversight, and with what obligations for communication, rollback or re-appraisal (CADTH, 2022; Bélisle-Pipon et al., 2021). This is conceptually similar to the challenge NICE faces when evolving treatment pathways can undermine earlier appraisal assumptions: in both settings, maintaining decision relevance depends on recognising that value is conditional on an evolving context and on having practical mechanisms to detect when that context (or the intervention) has changed sufficiently to warrant review.

10. Poor specification, transparency gaps, and limited reproducibility

A central challenge for HTA of frontier AI is that it is often unclear exactly what is being evaluated. In many cases, the technology is not described in enough detail to allow confident interpretation of the evidence, replication of results, or assurance that the evidence will still apply once the system is used in routine practice. Because frontier AI systems can be set up and used in different ways, and can influence multiple parts of care delivery, poor or inconsistent description does more than reduce transparency: it makes it difficult for HTA to determine what has actually been appraised and whether the available evidence applies to how the technology is used in practice.

The NIHR Generative AI Horizon Scan highlights wide variation in how generative AI technologies are described, noting that limited clarity can make it difficult to understand what is being deployed and therefore which evidence is relevant (Lanyi et al., 2025). Farah et al. (2024) similarly identify complexity, data requirements and lack of transparency as practical barriers to HTA of AI-based medical devices, arguing that HTA methods need to adapt to address these constraints. For NICE, this means appraisal cannot rely on broad or generic product descriptions. Instead, it needs to be grounded in clear descriptions of how the system is set up and used, including configuration choices, data inputs, the role the system plays in clinical workflows, and governance arrangements, because these factors directly influence performance, safety and equity for frontier AI systems.

A further difficulty is that important parts of the system may sit outside the direct control of those deploying it. Reliance on third-party foundation models, proprietary software interfaces or externally managed components can become important sources of uncertainty if they are not transparent, because they limit independent scrutiny and make it harder to judge whether NHS use remains consistent with what was originally evaluated (Bélisle-Pipon et al., 2021; MHRA/FDA/Health Canada, 2024). Where this information is missing, NICE's ability to link recommendations to a clearly defined system as used in practice is weakened.

WHO guidance reinforces this point, emphasising that governments may need to require ongoing disclosure from developers and providers, including sufficient documentation, to support safe use during deployment (World Health Organization, 2024). For NICE, this supports treating system description and disclosure not as optional background material, but as relevant evidence requirements and, in some cases, as conditions that must be met before a technology can be appraised.

Findings from NIHR Horizon Scanning illustrate the practical consequences of poor specification. Reviews of AI-based technologies and algorithms found that clinical trial records and related descriptions often omit key details about the algorithms used, leaving many technologies poorly characterised even within formal development pipelines (Forsythe et al., 2024). Similar gaps appear in the NIHR Generative AI Horizon Scan, where many technologies claiming clinical applicability did not specify the generative AI model or algorithm used (Lanyi et al., 2025). For HTA, these gaps are not just academic concerns. They make it harder to replicate findings, limit confidence that results will transfer across settings, and weaken NICE's ability to ensure that recommendations continue to apply to what is actually deployed.

Taken together, these findings strengthen the case for NICE to require minimum, standardised information as a condition of appraisal. This would include clear statements about the type of model used, how different software components are combined, whether and how external information sources are used, the intended role of the system in care pathways, and the technical and organisational arrangements needed to maintain safe performance over time, such as monitoring, update governance and oversight.

A similar issue is already familiar in HTA when assessing evidence from poorly described interventions or data sources. Just as NICE treats real-world or observational evidence with caution when key details about populations, comparators or data collection are missing, weak description and limited transparency in frontier AI reduce confidence that evidence applies to the technology as used in practice. In both cases, clear specification is a basic requirement for credible appraisal rather than a secondary reporting detail.

11. The accuracy-to-impact gap and overemphasis on technical metrics

A key challenge for HTA of frontier AI is the gap between strong technical performance and real-world impact on care pathways. Frontier AI systems may demonstrate impressive accuracy or other technical metrics in isolation, yet fail to deliver meaningful value in practice, or deliver value that is not captured by conventional endpoints used in evaluation. In some cases, apparent technical gains may even be offset by downstream harms, costs or unintended consequences elsewhere in the system.

The LSE framework and the AI for IMPACTS synthesis emphasise that evaluation of professional-facing digital health and AI technologies should not stop at performance metrics, but must also address integration and interoperability, organisational change, acceptability and trust, training requirements, governance and monitoring arrangements, and longer-term system effects (van Kessel et al., 2025; Jacob et al., 2025). This is particularly important for frontier AI because general-purpose capability can appear compelling when assessed in isolation, while real-world impact is determined by how the system is implemented and used in practice. Outcomes depend on workflow embedding, how clinicians interpret and act on outputs, local capacity and constraints, and the design of oversight and escalation processes. Moreover, system-wide and knock-on effects are an expected feature of frontier AI deployment, rather than an exception.

As a result, frontier AI can generate trade-offs across the care pathway, where improvements in one outcome are offset by harms or costs elsewhere. For example, faster clinical throughput or reduced clinician time per case may be accompanied by increased downstream investigations, new classes of clinical error, or additional workload associated with oversight, review and correction. Such patterns have been observed when AI systems are introduced to replace human readers in breast screening. System-wide impacts may also be unevenly distributed, leading to differential effects across patient groups and potentially altering health inequalities.

For HTA, this means that appraisal cannot rely on single headline performance metrics, even when those metrics appear strong. Instead, assessment may need to be grounded in explicit benefit–harm reasoning based on realistic models of care pathways, capturing downstream effects, resource implications and distributional impacts over time (Boverhof et al., 2024; Di Bidino et al., 2024). This mirrors existing HTA challenges where improvements in intermediate or surrogate outcomes do not reliably translate into overall patient benefit, and where value can only be established by examining effects across the full pathway of care rather than at a single decision point.

An illustrative example of this challenge is professional-facing tools designed to change how clinical work is carried out, such as ambient clinical documentation and AI scribe systems. Van Kessel et al. (2025) emphasise that evaluating these tools requires more than checking technical accuracy. It also requires understanding how they affect workflows, whether clinicians accept and trust them, and how well they fit into everyday practice. For HTA, these tools illustrate a common frontier AI challenge: benefits are often indirect, arising through changes in productivity, clinician experience, or service capacity, rather than through immediate clinical outcomes. At the same time, harms may be subtle, such as missing or incorrect documentation that affects downstream care. Economic evaluation therefore needs to model productivity carefully, include the costs of quality assurance, training, and governance, and reflect uncertainty about whether time savings lead to patient benefit, reduced backlogs, or improved staff retention. Empirical studies report reduced documentation burden alongside continued need for human review and editing, showing that design and governance strongly influence overall value (Haberle et al., 2024; Duggan et al., 2025).

12. Comparator choice problems and incremental-effect ambiguity

A core requirement of HTA is to estimate incremental benefit and incremental cost relative to current practice. For frontier AI, this requirement is often difficult to meet because the comparators used in evaluations are poorly defined or misaligned with NHS decision needs, making it hard to interpret reported effects as meaningful incremental gains for the health system.

Common problems include evaluations based on retrospective datasets that do not reflect real-world clinical pathways or decision points; comparisons against other algorithms rather than the relevant standard of care; and studies conducted in service models or health systems outside the UK, limiting transferability to NHS pathways, workforce constraints and levels of digital maturity. As a result, even when studies report strong performance, it can remain unclear what the AI is better than, in what context, and at what cost.

The medical AI HTA literature highlights that conventional frameworks often under-specify AI-specific considerations and that assessment needs to be broadened and structured to support comparability across evaluations and jurisdictions. This includes more explicit handling of continuous learning, explainability, interoperability and organisational impacts, all of which influence how incremental effects arise in practice (Boverhof et al., 2024; Farah et al., 2024). NICE's Evidence Standards Framework reinforces this need by requiring structured comparison with current care. For frontier AI, however, these requirements are harder to satisfy because the intervention's effects depend heavily on local configuration and workflow, and because system-wide responses, such as changes in clinician behaviour, demand, downstream testing, and oversight workload, are often part of the mechanism through which impact is realised. Unless comparator definitions and evidentiary expectations are made explicit at the scoping stage, by linking them clearly to the intended pathway role, care setting and service model, assessments are more likely to rely on proxy comparators or surrogate endpoints that do not translate into robust estimates of incremental effects and costs for the NHS (NICE, 2022). This weakens the ability of HTA to support decisions about value for money and prioritisation.

The Canada watch list adds a further practical dimension to comparator choice. It notes that widely available general-purpose AI tools are sometimes already used by clinicians without formal sanction, training or governance from employers or regulators, meaning that the technology may exist in practice even in the absence of organisational approval (Canada’s Drug Agency, 2025). This creates a specific HTA challenge for NICE because guidance may be sought only after some informal use has already started. In that situation, NICE is not being asked to police unsanctioned use or enforce regulatory or employer requirements. Instead, guidance can make clear that recommended introduction depends on minimum governance arrangements being in place. This matters more when a tool is already being used informally because the practical choice for services is often not use versus no use, but whether to leave use to individual discretion or bring it within formal organisational approval and oversight. NICE guidance can support that change by setting out the conditions under which use is recommended. Recommendations may therefore need to cover not only cost-effectiveness versus formal alternatives, but also the minimum organisational arrangements that must be in place before services introduce the tool, such as defined and approved use cases, training and competence requirements, clear oversight and escalation routes, and local monitoring or audit, so that use becomes consistent and governed rather than informal and uneven.

A similar issue arises in HTA when technologies are compared against usual care that is poorly described or delivered differently across sites. In those cases, estimating incremental benefit depends on being clear about what current practice looks like in reality, how it varies, and which parts of care the new technology is replacing or adding to. Frontier AI creates a sharper version of this problem, not because the comparator is necessarily another AI, but because general-purpose tools can be adopted informally and used in different ways before appraisal, and their impact depends heavily on local configuration and workflow integration. As a result, usual care may already have altered by the time guidance is sought, and may differ across settings in ways that effect estimates of benefit, risk and resource use.

13. Generalisability, transferability, and context dependence

A persistent challenge for HTA of frontier AI is that evidence generated in one setting may not generalise reliably to others. Real-world performance is shaped not only by the underlying model, but by how the system is configured and deployed in practice. Differences in population characteristics and case-mix, data availability and quality, interoperability with NHS systems, levels of digital maturity, workflow design, staffing and skill mix, training, and the design of oversight and escalation processes can all influence how the tool is used, what outputs it produces, and how those outputs affect clinical decisions.

These issues are amplified for frontier AI because such systems are frequently deployed as configurable services rather than fixed products, with prompts, retrieval sources, guardrails and workflow integration varying across sites. In addition, system-wide responses, such as changes in clinician behaviour, workload distribution or demand for downstream services, are often part of the mechanism through which impact is realised. Van Kessel et al. (2025) therefore emphasise that evaluation must consider integration and context, not only technical performance, and that professional-facing AI tools could be assessed in relation to how they are embedded in practice and interact with surrounding systems.

For NICE, this does not introduce transferability as a new idea, UK/NHS relevance is already considered explicitly in assessments. NICE could make this more systematic by requiring a standard transferability template for frontier AI, stating the UK/NHS assumptions for the evaluated use case and how they will be tested using UK real-world evidence after deployment. Appraisal may examine how performance, costs and resource impacts are expected to vary across settings and populations, what implementation conditions are assumed in the evidence base, and what controls are in place to detect and manage variation after deployment. Where evidence does not specify integration assumptions or treats implementation as trivial, conclusions about value may be misleading, because for frontier AI the degree of integration and governance often determines both achievable benefits and full lifecycle costs (van Kessel et al., 2025; Farah et al., 2024).

For NICE, this does not introduce transferability as a new idea, UK/NHS relevance is already considered explicitly in assessments. The additional point for frontier AI is that transferability may need to be handled more systematically because performance, costs and service impacts can vary substantially with local data, workflow integration, and governance. Appraisal may therefore make explicit what implementation conditions are assumed in the evidence (including integration and operating model), how outcomes and resource impacts are expected to vary across settings and populations, and what controls are in place to detect and manage variation after deployment. Where evidence is vague about integration assumptions or treats implementation as trivial, conclusions about value can be misleading, because for frontier AI the degree of integration and governance often determines both achievable benefits and full lifecycle costs (van Kessel et al., 2025; Farah et al., 2024).

A comparable challenge is encountered in HTA when evidence from trials conducted in specialist centres or highly controlled environments is applied to routine care. In those cases, NICE routinely considers whether differences in infrastructure, staffing or patient populations could alter effectiveness or costs. Frontier AI raises the same concern, but more sharply, because local configuration and governance are not peripheral details but core determinants of how the technology performs in practice.

14. Sociotechnical risk, human–AI team performance, and behavioral effects

In this report, sociotechnical means the benefits and risks of a technology depend not only on its technical performance, but on how it is used within real services, who uses it, in what workflow, with what training, oversight, incentives, and local governance. This is not unique to AI (many tests and HealthTech interventions have implementation effects), but it is especially important for frontier AI because it can shape clinical judgement and behaviour across multiple steps of care, and its impact can change as people adapt their practice around it. Therefore, a key challenge for HTA of frontier AI is therefore that outcomes depend on the performance of the human–AI system in use, not just on accuracy in technical evaluations. This can create risks such as clinicians placing too much trust in AI outputs, relying on them even when they are wrong, losing skills over time, or allowing errors to carry through later stages of care. Bélisle-Pipon et al. (2021) argue that AI introduces distinctive challenges for HTA because outcomes are shaped by these interactions between people and technology, meaning that ethical, social and behavioural effects are central to assessment rather than secondary concerns.

This makes transparency directly relevant to decision-making. The trilateral transparency principles stress that users need clear information about what AI systems are intended to do, what their limits are, and how they change over time in order to use them safely alongside human judgement (MHRA/FDA/Health Canada, 2024). Together, these perspectives highlight the importance of evaluating how AI systems are used in realistic clinical settings. They also point to a limitation of studies that rely only on retrospective data or technical testing: some risks only become visible once clinicians adapt their behaviour, develop reliance patterns, or create workarounds during routine use.

These concerns become stronger as frontier AI systems take on more complex or semi-autonomous roles, and as services adopt different models of human oversight. The same tool can be used with a clinician actively reviewing each output, with periodic supervision, or with the AI operating more independently with escalation only in defined circumstances, and these configurations change both safety risk and efficiency impacts in ways that may be relevant for HTA.In such cases, HTA is no longer focused on a single output, such as a risk score, but on a system that supports several steps of a clinical process. Frontier AI may, for example, collect patient information through conversation, bring together information from different sources, suggest possible diagnoses, recommend tests or treatments, and draft clinical notes. Recent benchmarking of an autonomous AI system in a real-world urgent-care setting shows both the potential and the assessment challenge: even when AI recommendations broadly match those made by clinicians, HTA must still consider rare but serious errors, performance in uncertain situations, and how trust, supervision and escalation processes affect safety (Hayat et al., 2025).

Frontier AI also raises additional governance concerns around how data are used. Records of interactions with the system, including prompts and outputs, often contain sensitive clinical information and may be reused for system improvement, analysis by suppliers or retraining. Governance discussions warn that without appropriate safeguards, these practices can undermine privacy and public trust, and that consent arrangements may need to be more detailed when systems are updated using data generated during routine care rather than data collected for research or development (Azad, 2025). This is particularly relevant for frontier AI because it may be used across many NHS tasks, increasing both the amount and sensitivity of data involved.

For HTA, governance matters because it affects whether a technology can be used safely and whether it delivers benefits in day-to-day care. By governance we mean the practical arrangements around use, how much the system can do on its own, how clinicians supervise it, when issues are escalated, and what training and accountability are in place, and where it affects whether the technology can be used at all, how patient data are handled, including consent and how prompts, outputs and logs are stored and protected. Poor arrangements can encourage informal workarounds, reduce trust and uptake, or lead to withdrawal after safety incidents. Overly complex arrangements can also reduce value by adding cost and slowing implementation. NICE can reflect this without trying to score weak governance in the cost-effectiveness model, or modelling different forms of goverance. The key point is that these governance, oversight and consent arrangements are conditions of safe use, and if they are not in place the technology should not be introduced. NICE therefore has an interest in making clear that these requirements must be in place for recommended use, by checking them explicitly in guidance and appraisal discussions and, where they involve meaningful time or cost (for example, training, supervision, audit/monitoring, local approvals and data-handling processes) ensuring these reflected in costs. Therefore, in practice, this is less about modelling the consequences of weak governance and more about specifying the minimum arrangements required for recommended use (Azad, 2025; World Health Organization, 2024).

A similar issue already arises in HTA for screening programmes, where the effectiveness of the test alone is not sufficient to determine value. Even when a screening test performs well technically, outcomes depend on professional behaviour and service organisation, including appropriate interpretation of results, adherence to referral and follow-up pathways, communication with patients, and downstream capacity for diagnosis and treatment. HTA therefore evaluates screening as an organised service rather than a standalone technology. Frontier AI raises the same type of challenge, because its real-world impact depends on how clinicians use and act on its outputs, and on how services are designed to support safe and appropriate use, rather than on technical performance alone.

An illustration of this issue comes from clinical reasoning and question-answering systems. An AI system (called Med-PaLM) showed design intention support expert-level medical reasoning in real clinical workflows (Singhal et al., 2025). From an HTA perspective, this is not a single diagnostic test, but a flexible reasoning capability. Its safety depends on clearly defined intended use, careful interaction design, and explicit limits on how outputs influence decisions. Evidence that focuses only on model answers risks overstating benefit if it ignores how clinicians interpret and act on outputs under time pressure and uncertainty. It may also understate harms caused by confident but incorrect responses. This reinforces the need for human-factors evaluation, clear system-in-use definitions, and monitoring over time.

15. Ethical, legal, and social issues are important domains to consider

A defining challenge for HTA of frontier AI is that ethical, legal and social considerations directly shape clinical impact and cannot be treated as peripheral or purely narrative concerns. When AI systems influence, shape or partially automate clinical decisions, issues such as privacy, consent, accountability, disclosure, equity and benefit–harm trade-offs become integral to whether the technology delivers value safely and acceptably in practice. These concerns are particularly acute in stewardship-like contexts, where decisions informed by AI can have consequences beyond the individual patient.

Ethical analyses in areas such as infectious diagnostics and antimicrobial stewardship emphasise the combined importance of transparency, privacy, fairness, accountability and meaningful human oversight, and argue that strong technical performance does not remove the ethical requirement for contestability and responsible governance (Panda, 2025). Di Bidino et al. (2024) similarly report expert consensus that AI-specific domains are critical for assessment and that core HTA models alone are insufficient, with ethics and ethical benefit–harm analysis ranked alongside accuracy, patient safety and data bias. For frontier AI, these domains are important because they condition how the system is used, trusted and relied upon in real clinical settings.
Ethical and legal considerations are intensified further because frontier AI systems can generate summaries and recommendations that appear authoritative, shaping clinician trust, acceptability and behaviour. This perceived authority sharpens questions of accountability: responsibility for acting on, supervising or contesting AI outputs may be distributed across multiple actors, including vendors supplying the AI-enabled service, foundation model developers, platform or cloud providers that configure or mediate access to the system, and NHS organisations that deploy it. This diffusion of responsibility complicates both governance and attribution of harms, and directly affects how safely the system operates in practice.

Guidance from World Health Organization on large multi-modal models reinforces the decision relevance of informed consent and disclosure, emphasising that patients should be made aware of material risks associated with AI use and that health workers should be trained on data protection risks, including those arising when protected health information is entered into AI systems (World Health Organization, 2024). For NICE, the implication is not that it should tell services exactly how to obtain consent. Rather, appraisal should be clear about what patients are told and how this is handled in practice, including who explains the AI’s role and risks, what training and safeguards staff need, and what arrangements are assumed for safe use, particularly where frontier AI affects clinical communication, decision support or documentation. Appraisal should also recognise any associated resource implications, such as staff time for explanation, training, supervision, or information governance. This is because weak or inconsistent practices in these areas can shape user behaviour and trust, increasing the likelihood of informal, unintended or unsafe use, and thereby changing how the technology functions in practice and what is effectively being appraised (van Kessel et al., 2025).

A comparable issue already arises in HTA when the effectiveness and safety of an intervention depend on consent processes, professional accountability and public trust, as is the case for screening programmes or population-level public health interventions. In these settings, ethical design choices are not ancillary considerations: they directly influence uptake, patterns of use and, ultimately, health outcomes. Frontier AI presents the same underlying HTA challenge, but in a more acute form, because governance, consent and accountability arrangements shape day-to-day use of the technology and therefore determine how benefits and harms materialise in practice.

An illustrative example of this challenge is generative AI integrated into electronic health record workflows designed to draft patient communications provides another frontier-adjacent case. Early empirical evidence suggests promising performance in some settings, while also highlighting the need for careful evaluation as use increases and systems evolve (Hu et al., 2025) due to questions about equity, acceptability, trust, and accountability, as well as the need for clear rules about review, editing, and final responsibility. Benefits and harms are often shaped by organisational policies and clinician behaviour, supporting the case for NICE to include expectations around human–AI team performance, transparency, and post-deployment monitoring in appraisal and recommendations.

16. Organisational readiness, implementation burden, and adoption barriers

A key challenge for HTA of frontier AI is that showing clinical promise does not, on its own, mean a technology can be implemented or sustained in routine practice. Whether AI is adopted depends heavily on organisational readiness, including procurement routes, governance capacity, staff training, changes to workflows, and the availability and quality of supporting data and digital infrastructure. Industry and system-level evidence points to persistent barriers in adoption, procurement and funding, alongside data and interoperability constraints, which together determine whether digital and AI-enabled technologies move beyond pilots and early adopters into everyday care (ABHI, 2023; ABHI, 2025).

Frontier AI increases these challenges by placing a much greater burden on implementation. Safe and effective use often requires ongoing monitoring, active change management and coordination across clinical, technical, information governance and operational teams. Many organisations do not yet have these capabilities at the scale required. In addition, practical constraints linked to procurement processes, cybersecurity requirements and infrastructure capacity can mean that technologies which appear cost-effective at appraisal are difficult to implement or maintain in real-world settings. AXREM highlights these issues in relation to imaging AI and digital infrastructure more broadly, pointing to the combined impact of procurement complexity, cyber risk management and assurance requirements (AXREM, 2025a; AXREM, 2025b).

These constraints are particularly important for frontier AI because routine operational tasks, such as managing cybersecurity risks, applying updates, or maintaining compliance, can directly affect regulated change control and the continued relevance of the evidence base. This creates practical tensions between security, assurance and ease of evaluation that need to be managed explicitly. Together, these factors can produce implementation problems, where a technology that appears valuable and workable in a controlled pilot cannot be scaled safely, consistently or affordably without substantial additional investment in organisational capability, governance arrangements and process redesign.

WHO guidance adds a practical dimension to this readiness challenge by recommending that health-care professionals are trained to understand how large multi-modal models work and their limitations, recognise appropriate uses and risks, avoid automation bias, communicate effectively with patients, and identify cybersecurity threats associated with AI use (World Health Organization, 2024). These expectations imply not just a training requirement, but a broader set of organisational capabilities. Where these capabilities are lacking or unevenly distributed across NHS providers, feasibility, safety and consistent realisation of benefits may be compromised, with implications for equity of access to frontier AI and for the long-term sustainability of its economic value.

A similar issue is already familiar in HTA when interventions require significant service redesign, workforce training or new governance arrangements to deliver their intended benefits. In such cases, cost-effectiveness depends heavily on assumptions about organisational capacity and implementation, and technologies that perform well in pilot settings may fail to scale without further investment. Frontier AI presents the same challenge in a more pronounced form, because implementation demands are ongoing rather than one-off, and organisational capability directly determines whether benefits can be achieved and maintained in routine care.

An example of this challenge is frontier AI in surgery. Reviews in musculoskeletal surgery describe a move towards systems that play a bigger role in planning procedures, guiding what happens during them, and shaping clinical decisions (Oettl et al., 2025). In these settings, the HTA challenge is not only whether the technology works, but whether it can be used safely in a high-risk environment and with clear responsibility for decisions. Appraisal may therefore need to consider how well the system has been tested, whether clinicians can understand and challenge its outputs, whether performance is reliable across different patient groups, and what ongoing checks are needed once it is in use. These factors also make it harder to define the intervention clearly and to estimate its full costs within standard HTA approaches.

17. Benchmarking, independent validation, and market comparability challenges

In crowded AI markets, the main HTA question is no longer whether AI can work at all, but which specific product and way of using it offers acceptable value, safety and equity for the NHS. This is already clear in areas such as imaging AI, where many technologies aim to perform similar tasks and the appraisal challenge is about comparison rather than simple effectiveness (Fasterholdt et al., 2022).

For frontier AI, this comparison problem is harder because products change quickly and do not exist as a single, fixed tool. How these systems perform depends not only

on the supplier or the underlying model, but also on how they are set up, how they fit into clinical workflows, and how clinicians use them in practice. As a result, differences in outcomes may reflect local implementation choices and context as much as differences between products themselves, making direct comparison more difficult.
Unlike more traditional technologies, frontier AI does not behave as a single, stable component. Its outputs are shaped by a combination of factors, such as the information it draws on, the rules that guide or constrain its outputs, the way it connects to other systems, and how it is embedded in everyday work. Because these elements can vary across organisations, both effectiveness and risk can differ substantially between sites, even for the same nominal product. This makes it harder to identify whether observed differences are due to product quality, local use, or ongoing changes over time.

Evidence from imaging AI shows that value assessment discussions already go beyond technical performance to include organisational and legal considerations, but high standard comparative assessments remain uncommon, with much of the literature relying on narrative or exploratory analysis (Fasterholdt et al., 2022). For frontier AI, this increases the risk that HTA conclusions are shaped by fragmented, supplier-led evidence that is difficult to compare, replicate or maintain as systems evolve. This poses a practical challenge for NICE in producing guidance that supports consistent commissioning decisions and remains relevant as technologies are updated and adopted across the NHS.

In response, frontier AI strengthens the case for more structured and standardised approaches to evidence assessment. This could include tailored evidence templates that clearly describe how systems are configured and intended to be used (Farah et al., 2024), and, where possible, independent benchmarking or reference standards that take account of real-world use as well as technical performance. Such approaches would help appraisal committees distinguish genuine differences in value and risk from differences arising from study design, selective reporting or short-term implementation choices.

A similar issue already arises in HTA when several technologies compete within the same clinical area, and the available evidence is mixed or selectively reported. In those situations, clear comparators, consistent outcome measures and independent validation are essential for fair and reliable decision-making. Frontier AI presents the same challenge in a more pronounced form, because technologies evolve rapidly and because how they are set up and used can matter as much as who supplies them.

18. Disruptive pathways change and system-wide propagation

A key challenge for HTA of frontier AI is that its main impact may come not from small improvements to existing tasks, but from changes to how care is organised and delivered. Frontier AI can be disruptive not only because it may reduce costs or be easier to scale, but because it can be used with different mixes of staff, different levels of supervision, and in different care settings. In practice, this means it can change who delivers care, where care takes place, and how responsibilities are shared, rather than simply improving individual steps within existing pathways.

INAHTA notes that disruptive technologies need to be assessed in ways that go beyond clinical performance, to include organisational and economic effects, as well as ongoing monitoring using real-world evidence. It also emphasises that the way technologies are adopted creates its own costs and demands on resources, which could be made explicit rather than treated as implementation issues outside the appraisal process (INAHTA, 2022). For frontier AI, this is especially important because changes in deployment models, supervision arrangements or workforce roles may be central to the benefits being claimed.

These challenges are greater for frontier AI because it is often designed to be used across multiple tasks and services, making changes across several care pathways more likely than with earlier digital health technologies. As a result, the HTA task is not just to assess gains within a single pathway, but to understand how changes in one part of the system affect other services, staff roles and use of resources. Improvements in one area may lead to increased demand, workload or costs elsewhere, which narrowly focused evaluations may miss.

For NICE, this means that careful scoping, proportionate use of conditional recommendations and ongoing review of guidance are particularly important when frontier AI is expected to change how care is delivered over time. A similar issue already arises in HTA when new technologies or service models change care between settings or change who provides it, making it insufficient to assess value within a single pathway alone. Frontier AI brings this issue into sharper focus because it has the potential to reshape several parts of the system at the same time.

An example case of this challenge concerns agentic or near-autonomous systems used in virtual urgent care or autonomous screening. In these settings, the comparison is no longer simply between a clinician and an algorithm. Evaluation must consider how care pathways are redesigned, how cases are escalated, and how safety is managed when presentations are unclear or conflicting. Even when average agreement with clinicians is high, HTA must still examine the risk of rare but severe errors, whether error categories are well understood, how well performance generalises across populations, and how governance arrangements manage risk as systems change (Hayat et al., 2025; Ahmed et al., 2025). These cases show why comparator choice often needs to be made at the pathway level, and why economic evaluation must include downstream service effects, capacity impacts, and governance costs, not accuracy alone.

19. Additional frontier-AI-specific challenges that require explicit handling in HTA

Alongside the core challenges outlined above, frontier AI also raises a set of additional issues that affect how HTA decisions operate in practice. These issues are not as central as the above core methodological challenges, but they can still influence whether decisions are applied consistently, can be implemented safely, and remain appropriate over time.

19.1. Assessor capability and multidisciplinary evaluation at pace

Assessing frontier AI often requires input from several areas of expertise, including clinical practice, digital systems, safety and risk management, and how people and organisations actually use technology. When access to this mix of expertise is limited, or when multidisciplinary input is applied inconsistently, assessments can take longer and conclusions may vary across HTA programmes. This creates uncertainty for innovators and commissioners and can undermine consistency in decision-making (Fasterholdt et al., 2022; Farah et al., 2024).

This is not just a matter of individual skills, but of assessment process. Without clear and reliable ways to bring different perspectives together, evidence expectations may change between appraisals, making it harder to judge what is required and harder to compare decisions over time.

A similar issue already arises in HTA for complex interventions, such as service redesigns or diagnostic pathways, where robust assessment depends on coordinated input from clinical, economic and organisational experts. In those cases, the quality and consistency of decisions depend as much on how multidisciplinary input is organised as on the evidence itself.

19.2. Fragmentation of assessment frameworks and inconsistent evidence expectations

The growing number of AI-specific assessment frameworks, together with inconsistent ways of turning these frameworks into concrete evidence requirements, creates practical problems for those assessing the evidence. Developers may be unsure which evidence is expected, while assessors may struggle to decide which AI based framework to use. This reduces consistency across evidence appraisal findings and decisions.

For NICE, this strengthens the case for a clearly defined and standardised AI annex requirement, setting out explicitly what evidence is needed and how it should be presented in relation to key decision questions. This would help reduce unnecessary variation and uncertainty, and improve comparability across appraisals (Farah et al., 2024; Jacob et al., 2025).

A similar challenge already arises in HTA when evidence appraisals use different outcome measures, modelling approaches or reporting formats. In those cases, NICE often standardises requirements to enable fair comparison and consistent decision-making. Frontier AI raises the same issue, because without common structures, it becomes difficult to judge relative value across technologies.

19.3. Deployment governance burden and reassessment in practice

Frontier AI typically requires ongoing governance once it is in use, rather than only checks at the point of initial approval. Experience with imaging AI shows that safe and effective deployment often depends on structured approaches such as phased roll-out, clear implementation guidance, multidisciplinary oversight, and regular review of performance. These measures help ensure that the technology continues to perform as expected and that risks are identified and managed as use increases (Fasterholdt et al., 2022; AXREM, 2025a). For frontier AI, these governance activities are not optional add-ons but routine parts of using the technology at scale. They affect whether a system can be implemented safely, how much it costs to run, and whether benefits can be sustained over time. For HTA, this means they could be considered as part of appraisal rather than treated as issues to be dealt with after a recommendation is made.

A similar situation arises in HTA for technologies that are commissioned and paid for as ongoing services rather than one-off products. In those cases, the value of the technology depends on continued quality assurance, oversight and review, rather than solely on the evidence available at the time of initial assessment.

19.4. Data sovereignty, cross-border services and jurisdictional governance

Data governance becomes especially important where frontier AI relies on foundation models, cloud services or vendor-hosted systems that store, process or log data outside the immediate control of the NHS. This can involve data moving between UK nations, or being processed in other countries. The Canada watch list highlights data governance and compliance across jurisdictions as key factors affecting feasibility and long-term trust in AI systems (Canada’s Drug Agency, 2025).

Where frontier AI depends on services that operate across organisational or national boundaries, the governance arrangements are not just a technical detail but part of the intervention itself. For NICE, these arrangements are relevant insofar as they may affect the feasibility, consistency, safety and equitable deployment of the technology within the health system, while broader questions of legal compliance and regulatory oversight may also sit with other bodies responsible for data governance and jurisdictional requirements.

A similar issue already arises in HTA when commissioning services that rely on external providers or cross-border supply chains, such as outsourced laboratory testing or specialist imaging services. In those cases, NICE would not usually assess contractual or governance arrangements in themselves, beyond considering whether any necessary regulatory approvals or implementation requirements are in place, and whether these have implications for feasibility, resource use or consistency of delivery that are relevant to the appraisal.

19.5. Environmental sustainability and system resilience

Environmental impacts are increasingly relevant background considerations for AI in health care. The Canada watch list highlights the environmental costs associated with training and operating large AI models, particularly where energy use and data storage increase as systems are adopted more widely (Canada's Drug Agency, 2025). OECD guidance also notes that uneven or poorly coordinated adoption can embed inefficiencies that persist over time (OECD, 2024).

Although environmental impacts are not core HTA outcomes, they can influence affordability, procurement constraints and the resilience of health systems, especially where operating costs scale with use. This suggests a need for proportionate consideration alongside other implementation factors, rather than full environmental assessment.

A comparable issue already arises in HTA when technologies place additional demands on infrastructure or supply chains. In those cases, considerations such as system resilience or long-term operating burden are not primary endpoints, but they can still affect feasibility and long-term value. Frontier AI raises a similar issue because environmental and resource demands may grow as adoption increases.

19.6. Liability, accountability and responsibility across the value chain

For frontier AI, responsibility for how the technology performs in practice is often shared across several actors, including the developers who build the system, platform or cloud providers who host or configure it, NHS organisations that deploy it, and clinicians who use it. WHO notes that this distribution of responsibility can make accountability and redress more difficult, particularly where patients are not aware that a large multi-modal model has been involved in their care (World Health Organization, 2024). The Canada watch list similarly highlights the risk that responsibility is placed mainly on individual clinicians or patients, rather than being clearly allocated at an organisational level, and calls for clearer policies and accountability arrangements (Canada’s Drug Agency, 2025).

For NICE, these issues matter because unclear accountability can increase the practical burden of governance, slow adoption, and affect safety and consistency of use. A similar challenge already arises in HTA for technologies where responsibility is shared across providers or settings, such as networked diagnostic services or outsourced testing. In those cases, how accountability is allocated directly affects both the risks of harm and the feasibility of safe implementation, and therefore needs to be considered as part of the appraisal rather than treated as an external legal issue.

19.7. Reliability of the evidence trail and multi-use capability

NIHR horizon scanning shows that AI innovation moves rapidly, while public documentation and trial records often lack the detail needed for robust comparison, increasing the risk of selective reporting and inconsistent evaluation across technologies (Forsythe et al., 2023; Forsythe et al., 2024; Lanyi et al., 2025). This problem is more pronounced for frontier AI because a single system can be used for multiple clinical purposes with little or no technical change. As a result, evidence generated for one use may not reliably apply to another, even though the underlying system appears unchanged. Different uses can carry different risks, affect different parts of the care pathway, and generate different costs and benefits. For HTA, this means that appraisal needs to be explicit not only about the technology being assessed and the configurations covered, but also about the specific use case to which the evidence relates. This overlaps with the challenge of configuration and specification, but the distinct issue here is whether evidence generated for one use can reasonably be applied to another when the underlying system appears unchanged. A similar issue already arises in HTA for technologies with multiple indications, where evidence must be linked to each specific use rather than assumed to generalise from broad or generic performance claims.

20. Health economic challenges specific to frontier AI appraisal

20.1. Why economic evidence for frontier AI is hard to interpret and compare

Frontier AI creates economic evaluation challenges that go beyond the already familiar difficulty of capturing implementation costs for digital health technologies. The central issue is that frontier AI often functions at the same time as a clinical tool, a shared platform and an evolving service. This makes it harder to define not only what benefits are delivered, but also how the technology is used over time, what it replaces, and what ongoing costs are required to keep it operating safely and effectively. These features are poorly reflected in simple, one-off pricing models.

CHEERS-AI helps explain why economic evidence for AI is often difficult to interpret and reproduce. It was developed in response to persistent gaps in reporting, particularly a lack of clarity about how AI systems operate and how they influence care pathways, factors that can change cost-effectiveness results (Elvidge et al., 2024). For NICE, this reinforces that economic evidence for frontier AI is not only a question of modelling technique. It also depends on clear, standardised descriptions of how the system is used in practice and what assumptions underpin the analysis, so that committees can understand what has actually been costed and valued (Elvidge et al., 2024).

Ghabri (2025) further notes that AI-based health technologies tend to evolve quickly, with frequent updates, learning effects over time, changing organisational impacts and dynamic pricing. Both the technology itself and the relevant comparators may therefore change during the period in which value is being assessed. This makes static economic models less appropriate. For frontier AI, this is especially important because many value claims relate to productivity, workflow change and service redesign rather than narrowly defined clinical outcomes, and these effects may emerge gradually rather than immediately (Ghabri, 2025).

Taken together, these features mean that the economic assessment of frontier AI faces a distinct set of challenges that are not well captured by conventional approaches. The key economic challenges for frontier AI are outlined below.

21. Key economic challenges for frontier AI

21.1. Unstable cost drivers and changing units of analysis

Frontier AI is often priced through subscriptions, enterprise licences, per-user fees or usage-based models rather than a fixed purchase price. Costs therefore depend on how much the system is used, which may increase as adoption grows, as the system is applied to new tasks, or as workflows change. As a result, both budget impact and cost-effectiveness can alter over time rather than remaining stable. A comparable issue is already familiar in HTA for medicines subject to patient access schemes or activity-based reimbursement. In those cases, the overall cost to the health system depends on how widely the medicine is prescribed and for how long, rather than on the headline price alone. Frontier AI presents the same challenge, because its economic impact is driven by patterns of use and uptake, not just by the nominal price of the technology.

21.2. Lifecycle-based cost structures and recurring assurance costs

Maintaining safe and effective use of frontier AI typically requires ongoing activities such as monitoring performance, detecting drift, auditing outputs, responding to incidents, managing cybersecurity risks, validating updates and retraining staff. These activities usually increase as deployment increases across services or sites. Transparency principles explicitly require lifecycle communication, update notifications and performance monitoring, all of which involve sustained organisational effort and therefore recurring costs (MHRA/FDA/Health Canada, 2024). Predetermined change-control approaches similarly imply repeated verification and impact assessment as systems evolve over time (FDA, 2025). For HTA, this means that the list price or licence fee of a frontier AI system rarely captures its full economic footprint. The costs of governance, assurance and ongoing oversight are part of what it takes to deliver the intervention safely and consistently, and could therefore be included in economic evaluation rather than treated as background implementation costs.

A comparable situation already exists in HTA for long-term therapies or implantable devices. In those cases, value assessment routinely includes the costs of follow-up appointments, monitoring tests, maintenance, and management of complications over time. Frontier AI raises the same type of issue: the benefits of the technology cannot be separated from the ongoing activities required to ensure it continues to perform as intended.

21.3. Implementation and adoption as cost-bearing interventions

INAHTA emphasises that disruptive technologies generate organisational and economic impacts, and that the strategies used to adopt and implement them also consume resources (INAHTA, 2022). For frontier AI, this includes investment in governance capacity, staff training, workflow redesign and ongoing monitoring. Treating these costs as outside the economic model risks overstating value for money. A comparable issue arises in HTA for service redesigns or complex public health interventions, where implementation costs can be decisive for cost-effectiveness.

21.4. Induced demand, pathway feedback and downstream costs

Frontier AI can increase downstream activity by identifying more cases earlier, altering referral patterns or lowering clinical thresholds for further investigation. This can change costs to earlier points in the pathway while benefits accrue later, or increase demand in services that already face capacity constraints. These effects require economic models that capture downstream resource use and feedback across care pathways, rather than assuming simple one-for-one substitution.

A similar issue is well recognised in HTA of screening and diagnostic expansion. In those settings, improved detection often leads to short-term increases in workload and costs, including additional tests, referrals and follow-up assessments. Once a condition is identified, patients may also enter longer periods of monitoring or surveillance for recurrence of the condition, generating ongoing costs even when the screening intervention delivers long-term health benefits through early detection of a condition. Frontier AI raises the same type of challenge, because earlier or more sensitive detection can increase downstream activity and resource use that must be accounted for when assessing value for money.

21.5. Linking staff time savings to outcomes and costs

Many frontier AI tools support clinicians rather than replacing them. Time savings may be redistributed to other tasks rather than converted into direct financial savings, and may not generate health gains unless they lead to measurable improvements such as reduced waiting times, lower staff burnout or improved retention. NICE's Evidence Standards Framework already anticipates the need to link resource impacts to outcomes; frontier AI makes this linkage especially important to specify explicitly rather than assume (NICE, 2022).

A similar issue is well recognised in HTA of digital workflow and efficiency tools. In those cases, evaluations often report time saved per task, but the economic value depends on how that time is actually used in practice. If time savings are spread thinly across staff or reinvested in unmet demand without reducing costs or improving outcomes, the claimed efficiency gains may not translate into cash-releasing savings or measurable health benefits. Frontier AI raises the same challenge, but at greater scale, because its use can affect multiple roles and activities across the care pathway.

21.6. Uncertainty, heterogeneity and structural change

The effects of frontier AI are likely to vary across organisations and patient groups, and frequent updates can make it uncertain whether benefits will persist over time. This is not a new problem for NICE, uncertainty is a feature of almost every appraisal, but frontier AI increases the likelihood that key parameters change during routine use (for example, as configurations change, users adapt, and systems are updated). As a result, economic conclusions may depend heavily on assumptions about how quickly the technology is adopted, how users learn to work with it, and whether organisational changes are sustained. This increases the importance of scenario and sensitivity analyses to explore how results change under different, plausible conditions, including explicit assumptions about the durability or waning of benefits over time. This is an issue NICE already grapples with for other technologies where long-term effects are uncertain and must be modelled, such as immunotherapies where treatment-effect waning is commonly assumed.

A comparable challenge also arises in HTA for interventions that rely on sustained changes in professional behaviour or service organisation. For example, early supported discharge programmes may show strong cost-effectiveness in sites with well-established community services and experienced multidisciplinary teams, but deliver smaller or delayed benefits in areas with limited capacity or high staff turnover. In those cases, outcomes depend on learning effects, local uptake and whether organisational changes are maintained over time. Frontier AI raises the same underlying issue, but with additional sources of instability because value may change as systems are updated and patterns of use evolve.

Heterogeneity is similarly familiar to NICE: treatment effects and costs often vary across patient subgroups and settings, but individualised effect estimation is typically infeasible. In practice, NICE relies on subgroup analyses where credible, and on scenario analyses and sensitivity analyses to explore plausible variation. The same approach can be applied to frontier AI, with heterogeneity explored both across patient groups (for example, language, multimorbidity, age, or protected characteristics where relevant) and across service contexts (for example, digitally mature versus less digitally mature settings, or high versus low baseline staffing capacity). Making heterogeneity explicit in this way when evaluating Frontier AI would help avoid over-reliance on average effects that may not reflect real-world deployment, and supports conclusions under plausible patterns of variation in performance, adoption, and sustained benefit.

21.7. Equity and affordability as economic as well as distributional issues

Frontier AI can deliver different levels of benefit and cost across settings, depending on factors such as digital maturity, organisational capacity and access to supporting infrastructure. As a result, average cost-effectiveness estimates may hide important differences in affordability and value between providers and populations. In practice, these differences can affect both equity of access and the long-term sustainability of adoption across the health system.

A similar challenge already arises in HTA when technologies depend on specialised infrastructure or workforce capacity that is unevenly distributed across providers. For example, interventions that require advanced imaging equipment, specialist interpretation, or dedicated digital systems may be cost-effective in large, well-resourced centres but much harder to implement affordably in smaller or less digitally mature trusts. In those settings, higher set-up costs, lower throughput and limited staff capacity can substantially reduce value for money or delay adoption altogether. Frontier AI raises the same issue, because its benefits and costs depend on local digital capability and organisational readiness, meaning that average cost-effectiveness estimates may not reflect the experience of all parts of the NHS.

22. What is distinctive about Frontier AI challenges, and what reflects established HTA challenges

Frontier AI poses a broad set of challenges for HTA, but these challenges are not evenly distinctive. Many resemble issues NICE already encounters when assessing complex interventions, diagnostics, digital health technologies and pathway redesign, where outcomes depend on implementation, workflow, organisational capacity, and behaviour rather than on a single fixed product. However, frontier AI sharpens and connects these familiar issues by directly undermining a core assumption that underpins NICE appraisal: that the intervention can be specified with sufficient clarity and stability for evidence generated during evaluation to remain valid for what is deployed, reimbursed and used in practice.

The most important differences for frontier AI arise from the likelihood of material change over time and the resulting risk that effectiveness, safety, equity, resource use and cost-effectiveness alter during routine use. The object of appraisal is not a fixed intervention but a configurable system whose behaviour depends on evolving combinations of model updates, retrieval sources, prompts and templates, guardrails, and workflow integration. This creates uncertainty that is qualitatively different from more uncertainty at baseline: it concerns what the technology is at any given moment, what it is permitted to do within workflows, whether evidence still applies after routine updates or reconfiguration, and whether observed outcomes can be attributed to the AI system versus the wider sociotechnical system around it. These features make intervention instability, evidence decay, and version ambiguity central HTA problems rather than peripheral complications.

Several specific frontier-AI challenges flow from this underlying instability. First, NICE faces a sharper problem of defining and binding the evaluable unit: the same nominal product may behave differently over time without an obvious version change, raising questions about what, precisely, has been appraised and recommended. Second, the durability of HTA conclusions becomes a core concern because evidence can become outdated as a normal feature of use; review timing therefore needs to be linked to the risk that system behaviour has changed, not only to fixed surveillance cycles. Third, frontier AI makes it harder to define significant change in a workable and proportionate way because meaningful changes can arise through multiple update routes that may be incremental and less visible. Finally, frontier AI's service-like nature intensifies post-deployment drift and accountability challenges: routine adjustments by vendors or deployers, limited public information about configurations, and (for continuous-learning systems) adaptation during use all increase the need to assess update mechanisms and lifecycle governance as part of the intervention, not just baseline performance.

By contrast, many other challenges identified in the section are strongly linked to challenges NICE already experiences, even if frontier AI makes them more acute. The gap between technical metrics and real-world impact mirrors familiar HTA problems where intermediate outcomes do not reliably translate into patient benefit and where system-wide effects and trade-offs matter. Comparator ambiguity and incremental-effect uncertainty reflect longstanding issues when usual care is variable, poorly characterised, or when evidence comes from retrospective datasets or non-UK contexts. Transferability and context dependence align with established concerns about applying evidence from specialist centres or controlled environments to routine care, though frontier AI heightens the importance of configuration, integration and governance as determinants of outcomes. Sociotechnical and behavioural effects resemble challenges already recognised in screening and other service-level interventions where professional behaviour, organisational design and downstream capacity shape outcomes, but frontier AI increases the salience because it can influence multiple steps of care and alter clinician behaviour in ways that are difficult to observe in technical testing.

Overall, the section supports a clear conclusion: frontier AI does not simply add more uncertainty to otherwise familiar appraisal problems. It makes change over time, and therefore the risk of evidence becoming misaligned with what is used in practice, the central distinguishing feature. This is the main reason frontier AI increases the need for appraisal approaches that can maintain assurance across the lifecycle, including clearer specification of the system-in-use, explicit handling of change thresholds, and greater reliance on monitored deployment and review where uncertainty cannot be resolved before adoption.

23. Recommendations to NICE for assessing frontier AI

The recommendations below are intended to be practical and usable. They focus on concrete improvements to scoping, evidence requirements and review triggers.

23.1. Specify precisely what the intervention is that is to be evaluated

To ensure that HTA conclusions apply to what is actually used in practice, NICE may require frontier AI modellers to define clearly and consistently the specific system configuration being assessed, rather than relying on high-level product descriptions or model names. This draws directly on how NICE already improves ease of evaluation in other hard-to-assess contexts: where evidence is complex or heterogeneous, the NICE Decision Support Unit has produced technical support documents that standardise what must be specified and how analyses should be presented, for example, a common modelling framework for pairwise and network meta-analysis (TSD 2), structured approaches for population-adjusted indirect comparisons such as MAIC/STC (TSD 18), and practical methods for using comparative observational data when randomised evidence is limited (TSD 17).

For frontier AI, NICE may need to operationalise the same principle to reduce ambiguity by standardising specification by requiring manufacturers to provide a structured system specification that makes the evaluable unit explicit, clearly defining what is being assessed and what a recommendation would apply to in practice. This specification could describe, in a consistent and auditable format, the components that together constitute the deployed technology, including: the core model(s) (the underlying AI model providing general capabilities); the orchestration layer (the logic that coordinates prompts, tools, rules and model calls); retrieval components (mechanisms that pull in external documents or data at run time); prompt templates (structured instructions or examples that shape model outputs); safety filters or guardrails (controls that limit unsafe or inappropriate outputs); human-in-the-loop design (where and how humans supervise, verify or override AI outputs); intended users; and integration points (how the system connects to electronic health records, workflows or other software). Together, these elements define the system-in-use rather than a notional algorithm in isolation and create a stable baseline specification that committees can interrogate, and that later evidence (including real-world evidence) can be mapped back to.

To reduce the risk that a general-purpose system is used beyond the assessed indication and role in the pathway, NICE could ask sponsors to be explicit about the system’s boundaries of use within the recommended care context i.e., what tasks, decisions or actions are out of scope for the recommendation, and to describe the safeguards that keep practice aligned with that scope. This is not about recommending use outside the indication; it is about making the limits of the recommended use clear and enforceable in day-to-day operation, where the same underlying model may be reconfigured or prompted in ways that change its function. Safeguards might include constrained workflows, restricted functionality, limits on downstream actions, or other controls that prevent the system being used for unevaluated tasks (Moor et al., 2023). This helps ensure that a NICE recommendation remains anchored to a bounded, evaluable intervention rather than drifting as the system is flexibly reconfigured in practice.

This system specification functions solely to define the object of appraisal. Requirements related to transparency, ongoing disclosure, monitoring, audit, or procurement-driven governance would be addressed elsewhere as decision-relevant evidence and lifecycle management considerations, rather than as part of the core intervention definition (MHRA/FDA/Health Canada, 2024; World Health Organization, 2024). In practice, it also strengthens NICE’s ability to apply its real-world evidence approach: the NICE real-world evidence framework is explicitly aimed at improving the quality and usefulness of RWE informing NICE guidance, and a clear system-in-use definition makes it much more feasible to plan, interpret, and act on RWE for rapidly updating technologies.

For example, an AI documentation assistant may be marketed as a single product, but its real-world impact depends on how it is configured and used. In one deployment, the system may generate draft clinical notes that clinicians must review, edit and sign, with no direct effect on orders, referrals or coding. In another deployment, the same underlying system may automatically populate structured fields in the electronic health record, suggest follow-up actions, or trigger downstream workflows unless actively overridden. These configurations differ in risk, workload, safety oversight and evidence requirements. A structured system specification makes clear which version has been evaluated and ensures that any NICE recommendation applies only to that defined mode of use, rather than to the technology in general.

23.2. Bind recommendations to version, configuration, and scope, with defined review triggers

To ensure that HTA conclusions apply to what is actually used in practice, NICE could require frontier AI appraisers to define clearly and consistently the specific system version, configuration and scope of use being assessed, rather than relying on product names or general descriptions of capability. This applies regardless of who produced the evidence, since the key point is that the version, setup, and intended use of the system must be defined in enough detail for the findings to be interpreted properly. NICE recommendations may then need to be linked explicitly to that defined version, setup, and use case, so that guidance reflects the intervention supported by the evidence, rather than a technology that may later be updated or used differently in practice.

Because frontier AI can evolve after appraisal, NICE may need a practical way to decide when a recommendation should be revisited. In most cases, this could be done by setting a small number of clear triggers for review, based on changes that could affect whether the recommendation still holds. These should focus on changes to scope, such as the intended indication, target population, role in the care pathway or level of autonomy, as well as clear signals of performance drift or safety concerns emerging in real-world use. For higher-impact applications, NICE could use existing time-limited or managed evidence approaches, for example by setting an early review point after a defined period of NHS use, alongside agreed plans for evidence maintenance and monitoring. The purpose is not to create a separate new process, but to make clear from the outset when NICE expects evidence to be revisited and what information will inform that review. This approach is consistent with NICE's existing recognition that digital technologies may require planned evidence maintenance rather than reliance on a single, static assessment (NICE, 2022). To make these triggers decision-relevant and linked to existing NICE processes, NICE can draw on its real-world evidence framework by ensuring that arrangements for post-deployment monitoring are set out in advance, in a way that can inform later review, similar to early use health technology assessments and consistent with NICE's existing early-use and evidence-generation approaches: clear questions, outcomes, data sources, analytic approach, and reporting that allow observed changes to be interpreted as evidence about value, safety, or equity (NICE, 2022).

To avoid uncertainty about when reassessment is needed, NICE could expect the evidence considered in appraisal to set out clearly which changes should prompt review, and who would be expected to notify NICE or trigger that review where those changes are identified. These triggers could include changes to the system itself, such as updates to the model, prompts, data sources, safeguards or integrations, and changes in how it is used in practice, such as a different target population, care setting, role in the pathway, or level of autonomy. This responds to concerns that thresholds for reassessment are often unclear for adaptive and agentic AI, and that leaving such decisions entirely to vendors or local services can weaken assurance, while overly frequent review can create unnecessary burden (Aquino et al., 2024). NICE can also encourage clear contractual arrangements across the value chain so that responsibility for monitoring changes, maintaining version control, and initiating review is clearly assigned, recognising that manufacturers may not always know when local use has changed in practice (Aquino et al., 2024).

For example, an AI triage system in urgent or primary care may initially be evaluated as a decision-support tool that flags higher-risk patients for clinician review, without directly influencing booking or referral decisions. The same system might later be reconfigured to automatically prioritise appointments, route patients to different services, or bypass clinician review for lower-risk cases. Although the product name and underlying model may be unchanged, these configurations differ substantially in their implications for access, workload, safety oversight and downstream resource use. Binding NICE guidance to the evaluated configuration, with clear triggers for review if the system's role or autonomy changes, and a guaranteed review point where warranted, helps ensure recommendations remain aligned with real-world use.

23.3. Require PCCP-like change governance plans or equivalent HTA change plans

To ensure that frontier AI systems remain safe, effective and evidence-aligned after approval, NICE may require appraisers to include a clear plan for how changes will be managed between formal HTA reviews. This plan should explain how updates are anticipated, tested, documented and communicated in routine use, rather than treating change as something that only triggers action at the point of reappraisal.

For frontier AI, this documentation can draw on the structure of the FDA's Predetermined Change Control Plan (PCCP), which sets out in advance what kinds of changes are expected, how they will be tested, what standards they must meet , and how changes will be documented and communicated (FDA, 2025). Where a formal PCCP is not in place, NICE could require an equivalent HTA change plan that includes explicit commitments to validation, record-keeping and notification when changes occur. The purpose would not be for NICE to approve the change plan itself as a separate assessment. Rather, it would give NICE and committees clearer sight of the types of larger changes that are expected after deployment, so they can decide at the outset what should trigger review and when reassessment may be needed. Therefore, this is not a recommendation for NICE to make regulatory judgements on safety, which sit with the MHRA and other assurance bodies. Rather, the aim is to ensure NICE has enough information to judge whether post-deployment changes could affect the continued relevance of the evidence and recommendation. This helps ensure that routine updates do not silently undermine the evidence base on which guidance depends.

Where appraisals involve ssystems that continue to change after deployment, for example because the model is updated using new data, feedback, or other information from routine use, NICE can also require a clear and bounded description of how learning occurs in practice. The purpose is not for NICE to judge whether that learning approach is technically appropriate in itself, but to understand whether these post-deployment changes could alter the intervention in ways that matter for the recommendation, and what this means for review. This could include what data the system is allowed to learn from, how often updates may occur, and what safeguards are in place to prevent performance drift or unequal effects across patient groups. The change plan can describe how learning-related updates are checked before release and how users are informed when behaviour may change (CADTH, 2022). Where update impacts are evaluated using observational or routinely collected comparative data, NICE can make expectations more consistent by explicitly anchoring analytic approaches to methods NICE already recognises for comparative real-world studies (for example, approaches reflected in the NICE Decision Support Unit’s methods guidance on observational comparative effectiveness).

For frontier AI deployed at scale or updated frequently, NICE may also expect proportionate arrangements for independent audit and impact assessment. WHO guidance recommends post-release auditing and independent evaluation for large multi-modal models, with transparent reporting of safety and performance signals over time (World Health Organization, 2024). NICE can translate this into evidence-maintenance conditions, for example by linking continued coverage of high-impact systems to periodic independent review of update effects and real-world outcomes.
For example, a clinical triage or documentation system that is updated monthly through a cloud service may undergo small but frequent changes to prompts, data sources or safety filters. Individually, these updates may appear low risk, but over time they can alter accuracy, clinician reliance or workload. A PCCP-like HTA change plan would make clear in advance which updates are expected, how each will be tested, when NICE or commissioners will be notified, and when accumulated changes should trigger closer scrutiny or re-review. This allows innovation to continue while maintaining confidence that the system being used remains aligned with what was originally assessed.

23.4. Strengthen transparency requirements

To support clear appraisal, consistent commissioning and safe use in practice, NICE could require frontier AI appraisals to include standardised transparency materials that translate the evaluated system configuration into clear information. This is not intended to replace or go beyond the regulatory and technical documentation that already defines intended use. Rather, the aim is to ensure that the key points most relevant to appraisal and implementation are presented in a clear, consistent, decision-facing form. These materials should not redefine the technology or control how it changes over time. Instead, they should make explicit what the assessed system is intended to do, how it should be used, and where its limits lie. This can build directly on NICE's established expectations for digital health technologies, where the Evidence Standards Framework emphasises that developers should provide clear information to support appropriate use, including what the technology does, how it should be used in practice, and what limitations or risks users and services need to understand (NICE 2022).

For frontier AI, clear information is especially important because the same system may appear more capable than the specific use case that has actually been assessed, and its performance can vary depending on how it is set up and used. Without a clear summary of the evaluated use case, there is a risk that the system is used beyond the supporting evidence or implemented differently across settings. NICE could therefore ask for a consistent summary of the key information already set out in technical and regulatory documentation, so that the points most relevant to appraisal and implementation are easy to identify and use. This could cover the intended purpose and role in the care pathway, known limitations, evidence gaps and subgroup performance, and how uncertainty is communicated to users (MHRA/FDA/Health Canada, 2024). The aim is not for NICE to impose a separate or higher transparency standard, but to make sure the information needed to define scope, interpret evidence, and frame recommendations is presented clearly and consistently.

WHO guidance reinforces this need by emphasising that large multi-modal models should not be used where mismatches between training data, language, population or context make safe and effective use unlikely (World Health Organization, 2024). NICE can operationalise this by ensuring that appraisal makes clear the settings and conditions in which the available evidence is likely to apply, and where important uncertainty remains, including assumptions about patient populations, care settings, workflows and data inputs. This does not depend on manufacturers stating where their own intended use is unsupported. Rather, it means that the limits of what can reasonably be concluded from the evidence are made explicit for committees to help them judge transferability and help services avoid inappropriate deployment.

To ensure transparency is meaningful rather than aspirational, NICE could also require consistency between the system description used for HTA and descriptions found in trial registries, marketing materials and deployment documentation, with clear explanation of any differences. NIHR horizon scanning shows that key technical details are often missing or inconsistent across public sources, even for technologies in formal development pipelines (Forsythe et al., 2024; Lanyi et al., 2025). Requiring a single, authoritative, system description, or a justified explanation where simplification is unavoidable, would reduce ambiguity for appraisal committees and support consistent application of version- and scope-bound NICE recommendations.

For example, a frontier AI system built on a general-purpose language model may be marketed as an AI clinical assistant or decision-support tool, suggesting broad clinical capability. However, the configuration evaluated by NICE may be limited to generating encounter summaries and highlighting relevant guideline excerpts, with no authority to recommend treatments, propose investigations or trigger actions in the electronic health record. Clear transparency documentation would make this distinction explicit, set out known limitations (such as reduced reliability in certain specialties, patient groups or languages), and state clearly that the system must not be used to generate treatment plans, prioritise diagnoses or initiate orders. This ensures that NHS organisations deploy the system in the specific, bounded way that was evaluated, rather than extending its use based on generic claims about the underlying model's capabilities.

23.5. Add a short frontier AI checklist to NICE guidance to ensure consistent reporting.

To support consistent appraisal of frontier AI, NICE may need to introduce a dedicated HTA annex that defines the domains that must be addressed for this class of technology. This would reduce reliance on ad hoc extensions of existing digital health frameworks and ensure that frontier-AI-specific risks and value drivers are considered systematically across topics.

Frontier AI can influence multiple parts of care delivery and evolve during use. Therefore, the annex could set out a core set of issues that should be considered in appraisal, even where detailed assurance sits elsewhere. These include bias and representativeness, the information needed for safe and appropriate use in practice, arrangements for monitoring changes over time, whether relevant external approvals are in place for areas such as cybersecurity and data governance, how responsibilities are shared across organisations, and any workforce training needed (Di Bidino et al., 2024). The aim is not for NICE to assess all of these areas in full or duplicate checks already carried out through legislation, regulation, or existing frameworks. Rather, the annex would help make clear which issues NICE should consider directly, which should be supported by existing external approvals or local organisational arrangements, and which conditions should be in place for NHS use. It should also set out the minimum information needed for appraisal, so committees can understand what the technology is, how it has been tested, how it is expected to be used, and what practical assumptions its use depends on. This would make expectations clearer, reduce unnecessary duplication, and help recommendations focus on whether the technology can be used safely, appropriately, and with good value in practice. Where important uncertainty remains, the annex could also indicate when a more cautious route is needed, for example linking use to further evidence generation and an earlier planned review. More consistent expectations across frontier AI appraisals would also reduce variation in evidence requests for similar technologies and support more predictable, consistent decision-making, while preserving committee judgement on what matters most in each case.

For economic evidence, NICE could require or strongly encourage use of CHEERS-AI when reporting economic evaluations of AI interventions. CHEERS-AI extends standard economic reporting by requiring clearer description of the AI system’s role in the care pathway and the implementation assumptions that influence cost-effectiveness estimates (Elvidge et al., 2024). For frontier AI, this provides a practical safeguard, helping committees see exactly what has been costed, for which use case, and under what implementation conditions.

For example, two technologies may both be described as clinical copilots, but one may be limited to drafting notes and surfacing relevant information, while another actively influences triage decisions or follow-up actions. Without a structured annex and minimum reporting expectations, appraisals may emphasise different domains or omit key assumptions, making comparison difficult. A frontier AI annex would ensure that both appraisals cover the same core issues, such as how the system fits into the workflow, what arrangements are needed for safe use, and what impact it may have on services, so NICE can compare value and risk on a more consistent basis rather than relying on uneven or selective evidence.

23.6. Strengthen NICE's capacity to assess frontier AI and standardise multidisciplinary assessment processes.

To support consistent, timely and proportionate appraisal of frontier AI, NICE could strengthen its internal capability and standardise how multidisciplinary expertise and accountability considerations are incorporated into assessment. Frontier AI can effect clinical practice, workflows, safety and governance simultaneously, which may mean appraisal needs both broader expertise and clear, repeatable ways of identifying what assumptions are being made about oversight, monitoring and responsibility in order for the technology to be used safely in practice. To support this, NICE may need a team to provide frontier AI methods support across appraisal programmes. They could advise on scoping and comparator choice, support interpretation of complex or evolving evidence, maintain institutional memory as technologies update, and coordinate access to relevant expertise, including clinical, economic, data, safety and organisational perspectives. The literature consistently highlights that clearer methodological infrastructure and standardisation improve comparability and reduce variability in AI HTA decisions (Farah et al., 2024; Jacob et al., 2025).

As part of this process, NICE could expect appraisals to set out clearly who is expected to oversee the system in practice and what needs to happen if problems arise. For frontier AI systems that can change after deployment, it is important to know who is responsible for monitoring performance, raising concerns, and starting a review when needed. The point is not that NICE would check governance in the same way as a regulator, but that committees may need this information to judge whether the technology can be used safely and consistently in the way described. This helps address a common problem with adaptive AI, where responsibility can become unclear once systems are in use (Aquino et al., 2024).

For example, an AI system may be supplied by one organisation, hosted on a third-party platform and configured locally by NHS providers. A stronger NICE assessment process would not treat this simply as an implementation detail, but would make clear what assumptions are being made about who is expected to monitor performance, report problems and initiate review. This is not a recommendation for NICE to take on functions that sit with regulators or other assurance bodies, such as oversight of ongoing safety issues. Rather, the aim is to ensure that appraisal and recommendations are based on a clear understanding of the arrangements needed for safe use and evidence review over time.

23.7. Require the human–AI system and sociotechnical fit as a core appraisal object

To ensure that HTA reflects how frontier AI actually affects care, NICE could treat the combined human-AI system, rather than the AI component alone, as the intervention under appraisal. This is because, when frontier AI is used for decision support, documentation, triage or workflow coordination, outcomes depend as much on how people interact with the system as on the technical performance of the model itself. The human–AI system definition makes it feasible to specify what must be monitored in practice (for example, override rates, escalation failures, uneven use across groups, or workflow disruption) and what would trigger review.

Evidence appraisals could therefore address the design features that shape real-world use and risk. These include how outputs are presented to users; how uncertainty, limitations and confidence are communicated; what training and competence are required; how escalation and override work in practice; and what safeguards are in place to reduce automation bias, over-reliance or inappropriate use. These factors directly influence safety, effectiveness and value, and are highlighted both in transparency principles for machine-learning-enabled devices and in AI-HTA literature that identifies sociotechnical integration as a key driver of outcomes (MHRA/FDA/Health Canada, 2024; Bélisle-Pipon et al., 2021; van Kessel et al., 2025).

To make this appraisal-ready, NICE can require that claimed benefits (for example, time saved or errors avoided) are explicitly linked to the conditions of use that make them plausible, such as the level of human checking required, the training burden, and the expected changes in clinician behaviour, rather than assuming that technical accuracy translates automatically into real-world impact.

Practically, this is directly applicable to frontier AI claims that are often framed around workflow/process improvements. NICE can ask developers to (i) define the intermediate outcome precisely (e.g., net time saved per consultation after human checking); (ii) provide evidence for how that intermediate change alters downstream care or service outputs (e.g., additional capacity leading to shorter waits, earlier treatment initiation, fewer missed follow-ups, reduced adverse events); (iii) quantify how those downstream changes translate into patient outcomes/QALYs within an economic model; and (iv) propagate uncertainty across each link. NICE's guidance also recognises that, particularly in diagnostics and device-like evaluations, decision-making often depends on linked-evidence modelling, combining evidence from different sources to connect intermediate measures (e.g., accuracy statistics) to final patient outcomes, and it explicitly expects the links (e.g., diagnosis to treatment to final outcomes) to be specified and justified. Where that link depends on correlated outcomes or surrogate relationships, NICE Decision Support Unit methods guidance (e.g., multivariate/bivariate meta-analytic approaches for surrogate validation) provides a concrete, NICE-recognised way to quantify the surrogate-to-final relationship and its uncertainty for use in modelling (NICE DSU, 2019).

WHO guidance reinforces the importance of formal oversight for AI used in clinical decision-making, including mechanisms to ensure appropriate use and protect patient rights (World Health Organization, 2024). In the NHS context, NICE can apply this pragmatically by expecting that higher-impact frontier AI deployments have clear local oversight arrangements, defined escalation pathways, and documented responses to audit findings and safety incidents. The aim is not to add new layers of governance, but to ensure that responsibility for how humans and AI work together is explicit rather than unclear.

For frontier AI systems that are agentic or partially autonomous, meaning they can initiate actions or shape decisions with limited human prompting, NICE could require an explicit description of autonomy as part of the appraisal. This could set out which tasks are assistive versus autonomous, where human confirmation is required, how uncertainty is conveyed, and what safeguards exist to prevent unsafe outputs. Where systems operate across multiple steps of care, evidence should include structured analysis of potential errors and safety-critical testing under realistic conditions, including ambiguous cases and incomplete information. This reflects evidence that rare but serious harms may only become visible in such settings, even when average performance appears strong (Hayat et al., 2025).

For example, a frontier AI tool may summarise patient information and suggest possible next steps during an urgent-care encounter. If clinicians see suggestions without clear indicators of uncertainty, limits or required oversight, they may treat outputs as authoritative rather than advisory. An appraisal that focuses only on diagnostic accuracy would miss this risk. Treating the human–AI system as the intervention makes it possible to assess whether interface design, training and escalation processes are sufficient to support safe use, and whether the claimed benefits are likely to be realised in routine NHS practice.

23.8. Require transferability analysis and UK-relevant real-world evidence planning as a core appraisal object

To ensure that evidence for frontier AI applies to NHS practice, NICE could require a structured assessment of transferability as a core part of appraisal. This would set out how results from studies or pilots are expected to translate into real NHS settings, where important differences in context could affect performance, safety or costs, and what mitigations are required to support safe deployment at scale.

For frontier AI, transferability cannot be assumed. How well a system works often depends on local configuration, how it fits into clinical workflows, the data it draws on, and organisational capacity. Appraisals could therefore identify which aspects of performance or resource use are sensitive to local conditions, how study settings differ from expected NHS deployment, and what local checks, adaptations or safeguards are needed before or during implementation. This strengthens the appraisal focus on

NHS ease of deployment rather than treating implementation as a secondary consideration.

For higher-impact frontier AI, NICE could also consider requiring a UK-relevant real-world evidence plan as part of appraisal, potentially as a condition of a time-limited or managed access recommendation that is explicitly reviewed after a defined period of NHS use (for example, after a couple of years), analogous in principle to established managed access approaches such as the Cancer Drugs Fund. This plan could explain how real-world performance, use, and service impact will be followed after deployment, including which outcomes will be tracked, how important changes will be identified, and what information would support a later review. This reflects NICE's existing recognition that some digital technologies require evidence maintenance over time, rather than one-off evaluation (NICE, 2022; FDA, 2025). The aim is not for NICE to take on safety monitoring functions that sit with regulators or other assurance bodies. Rather, it is to ensure that appraisal is linked to a clear plan for evidence maintenance and review. That boundary is especially important because the MHRA's National Commission into the Regulation of AI in Healthcare is currently advising on a new UK regulatory framework for AI in healthcare, with recommendations due in 2026 The Canada watch list highlights that factors such as privacy and data security, liability and accountability, data quality and bias, and data sovereignty can strongly influence whether AI systems can be adopted safely and consistently (Canada's Drug Agency, 2025). NICE can address these issues by strengthening transferability requirements. Appraisals could explain how the system meets UK data governance expectations, what integrations and data flows are required, where data are stored and processed, and what organisational controls are needed to maintain privacy, security and accountability over time.

Treating these considerations as part of transferability, rather than as later implementation details, recognises that for frontier AI, real-world performance and risk are tightly linked to local data, infrastructure and organisational capability. It also aligns with proportionate evidence-generation expectations, so that where uncertainty is high, a managed access recommendation can specify what UK real-world evidence must be generated (framed using NICE's RWE expectations) and what triggers a reassessment.

For example, a frontier AI tool evaluated in a digitally mature hospital with well-integrated records and dedicated oversight staff may perform well and appear cost-effective. If the evidence depends on those conditions, NICE may need to make this explicit in the recommendation, for example by recommending use only in settings with the necessary record integration, oversight, or trained staff in place. If deployed in NHS settings with different data quality, fewer integration points or limited monitoring capacity, the same system may behave differently or impose higher costs and risks. A review of how well the evidence is likely to apply in different NHS settings would help identify these differences and clarify what evidence, safeguards, or local conditions are needed to support safe and effective NHS use, including what should be revisited at a defined review point after real-world NHS deployment.

23.9. Strengthen comparator expectations and enforce explicit comparator justification at scoping

To ensure that value estimates for frontier AI are meaningful for NHS decision-making, NICE may need to apply its existing approach to comparator selection particularly carefully at scoping. Comparators are already defined through NICE's scoping process, but for frontier AI the rationale may need to be made more explicit where current practice is variable, informal, or altered by the technology itself.

For frontier AI, comparator selection is not a minor technical choice. These systems often change how work is done, redistribute tasks across staff groups, or reshape care pathways altogether. Appraisals could therefore make clear what is being compared with what, for example, standard care versus AI-supported practice, AI-supported practice versus existing digital tools, or a redesigned pathway versus the current service model. Making these distinctions explicit at scoping helps prevent comparisons that blur incremental effects or exaggerate benefits.

Where randomised trials are not feasible or appropriate, NICE could expect the use of well-justified alternative study designs. These could include a clear description of the assumed counterfactual, transparent methods for estimating what would have happened without the AI, and an explicit link between intermediate effects, such as changes in decisions, throughput or error rates, and downstream clinical, organisational and economic outcomes. This aligns with calls for evaluation approaches that recognise the practical constraints of assessing complex and adaptive AI systems (Boverhof et al., 2024; Farah et al., 2024). NICE can operationalise this by requiring that comparative real-world studies follow the expectations set out in its real-world evidence framework (for example, clear causal question specification, data source justification, and bias/confounding handling), and by drawing on DSU methods for comparative observational evidence (NICE, 2022; Faria et al., 2015).

By requiring comparator justification early, NICE can improve the clarity and consistency of frontier AI appraisals, reduce reliance on proxy or convenience comparators, and ensure that appraisal conclusions reflect real NHS choices rather than abstract technical benchmarks. Where uncertainty about the appropriate comparator (or about how standard care will evolve) is material at the point of appraisal, NICE can also connect comparator requirements to existing lifecycle tools, using time-limited or managed access-style arrangements to generate UK comparative evidence against the agreed NHS comparator and revisit the decision at a defined review point, analogous in principle to established managed access routes such as the Cancer Drugs Fund.

For example, a frontier AI triage tool might be evaluated against another algorithm or against clinician performance in a retrospective dataset. But the comparison that matters for an NHS decision may be different: how triage is actually carried out in practice by a clinical team, within existing workflows and staffing constraints. The point is not that comparator choice is missing from current health technology assessment, but that for frontier AI it may need to be framed more carefully where the technology is not simply replacing one existing tool, and instead is supporting staff or changing how the service works.

23.10. Take account of the real work and costs of putting frontier AI into practice.

To ensure that HTA conclusions reflect what can realistically be implemented in the NHS, NICE could treat procurement requirements, cybersecurity controls, infrastructure dependencies and assurance burdens as core inputs to appraisal rather than as secondary implementation issues. For frontier AI, these factors often determine whether a technology can be deployed safely, scaled across services and sustained over time, as well as its true cost to the system.

Frontier AI systems are frequently updated, tightly integrated with electronic records and other digital infrastructure, and subject to multiple layers of organisational and regulatory assurance. As a result, procurement constraints, interoperability requirements, cybersecurity obligations, update governance and assurance processes can affect feasibility and affordability. These are not incidental frictions: they shape whether a technology that appears cost-effective on paper can actually be used in practice, and at what ongoing cost. Where the evidence base relies on indirect comparisons across different settings or study populations, NICE can make expectations more consistent by explicitly aligning appraisals with methods NICE already relies on in other complex appraisals, such as DSU guidance on network meta-analysis and evidence synthesis (Dias et al., 2011) and on population-adjusted indirect comparisons (MAIC/STC) when trial populations differ (Phillippo et al., 2016). Evidence from AXREM highlights how duplicated assurance processes, overlapping cybersecurity requirements and fragmented procurement arrangements can impose substantial operational and financial burdens on providers, particularly where responsibilities are unclear or requirements differ across organisations (AXREM, 2025a; AXREM, 2025b). For frontier AI, such burdens can determine whether deployment is viable at all. NICE could therefore expect these constraints and costs to be made explicit in both economic analyses and implementation arguments, rather than treated as externalities outside the scope of appraisal.

OECD policy work reinforces that leaving governance and assurance expectations underspecified can itself create risk, including fragmented adoption, inequitable access and avoidable cost over time (OECD, 2024). For NICE, this supports clearer expectations within guidance about governance, assurance and interoperability for frontier AI, so that recommendations are grounded in realistic assumptions about how technologies will be procured, secured and maintained across the NHS.

Where value claims are based largely on productivity, NICE may need to consider more carefully how any reported time savings would translate into real benefits in NHS practice. This is unlikely to be something developers can define on their own, because any released staff time may be used differently across settings. Instead, appraisal and committee discussion may need to consider whether the time saving is likely to be meaningful in practice, how it could be used in different services, and whether it is likely to improve throughput, reduce delays, support other clinical activities, or lead to downstream benefits for patients. Making these pathways explicit avoids assuming that time savings automatically generate health gains or cash-releasing efficiencies, which is often not the case in practice.

In summary, economic appraisal could include ongoing costs linked to maintaining performance and safety, such as monitoring, audit and incident response, cybersecurity controls, validation of updates, and workforce training or retraining. Because frontier AI systems are frequently updated and may be used more intensively over time, appraisals could also explore how costs and outcomes change under different assumptions about update frequency, monitoring intensity, growth in use, and learning effects (Ghabri, 2025; MHRA/FDA/Health Canada, 2024; FDA, 2025). Scenario-based modelling helps avoid relying on a single, static cost-effectiveness estimate that may not hold as systems evolve.

For example, a frontier AI tool may appear affordable based on licence fees alone, but require significant additional investment in cybersecurity controls, system integration, repeated assurance reviews and staff time to manage frequent updates. If these costs and constraints are not considered at appraisal, NHS organisations may struggle to implement the technology consistently or may abandon it after initial deployment. Treating procurement, infrastructure and assurance burdens as relevant helps ensure that NICE recommendations reflect the full economic and practical implications of adopting frontier AI in routine care.

23.11. Strengthen economic evidence to reflect lifecycle use and uncertainty for high-update frontier AI.

As explained above, the economic case for frontier AI depends not only on the technology itself, but on how it is used and maintained in practice over time. NICE could require economic appraisals to capture full lifecycle costs and changing patterns of use, rather than focusing mainly on upfront purchase or initial implementation. For frontier AI, value for money depends as much on what it costs to operate safely over time as on what it costs to acquire.

Given the uncertainty inherent in adaptive and configurable systems, NICE could also set minimum expectations for uncertainty analysis in economic models. This could include sensitivity analysis around key drivers such as governance intensity, update and monitoring requirements, and assumptions about uptake and sustained use. Where there is good reason to expect variation across sites or populations, models could also explore heterogeneity rather than relying solely on average effects.

Finally, to support transparency and reproducibility, economic evaluations could state clearly how the frontier AI system affects the care pathway and what integration and implementation assumptions underpin the analysis. CHEERS-AI highlights that lack of clarity in these areas is a common weakness in AI economic evaluations, making it difficult to judge what has actually been costed and valued (Elvidge et al., 2024). Addressing this ensures that economic conclusions correspond to the system that would be deployed in NHS practice, rather than to an abstract or idealised version.

For example, a frontier AI tool may be promoted as cost-effective because it reduces clinician documentation time. A lifecycle-based economic appraisal would show whether this saving persists as use scales, what additional costs arise from monitoring and update validation, and whether freed-up time leads to shorter waits, higher throughput or improved outcomes. This allows NICE to judge value for money based on realistic patterns of use and cost over time, rather than on optimistic assumptions at the point of purchase.

23.12. Use proportionate conditional recommendation pathways and maintained guidance for high-update frontier AI

Where frontier AI systems change frequently or have uncertain real-world effects at the point of appraisal, NICE could use proportionate conditional recommendation pathways that link access to ongoing evidence generation and review. This aligns with mechanisms NICE already uses for other technologies where uncertainty is central: managed access routes explicitly provide NHS access while additional evidence is generated to address key uncertainties, followed by a further NICE decision about routine commissioning. This allows timely use of promising technologies while recognising that their safety, effectiveness and value may only become clear through monitored deployment.

Policy work on frontier AI governance highlights transparency, independent evaluation and post-deployment monitoring as practical ways to manage rapidly evolving systems under uncertainty (Bommasani et al., 2025; DSIT, 2025). NICE can reflect these principles through managed access or similar arrangements, in which early or limited use is explicitly tied to defined monitoring requirements, data collection plans and scheduled review points. This approach acknowledges that for frontier AI, uncertainty is not always resolvable before deployment and must instead be managed over time. This differs from appraisal of more stable digital technologies, where functionality changes slowly and a single pre-deployment evaluation may be sufficient. For frontier AI, the HTA task is less about reaching a final judgement at one point in time and more about maintaining assurance as systems evolve. Conditional recommendations provide a structured way to balance access with oversight, helping to manage changing risk, performance and value without undermining safety, equity or long-term sustainability.

WHO guidance reinforces the importance of credible stop mechanisms, noting that deployers may have a responsibility to suspend or withdraw large multi-modal models if continued use risks harm, even where this is not legally mandated (World Health Organization, 2024). This would not mean NICE taking on a safety regulation role, which sits with the MHRA and other bodies. Instead, for high-impact frontier AI, recommendations may need to be clear about what would happen if important external concerns arise, for example if regulatory approval is withdrawn, a formal safety warning is issued, agreed monitoring is not carried out, or the conditions of use are not met. NICE can also encourage commissioners to ensure that contracts and governance arrangements make suspension or withdrawal feasible in practice. Explicit stop rules help ensure that conditional access remains a genuine risk-management tool rather than an automatic route to routine adoption.

For example, a frontier AI triage system might be recommended for limited use in specific services, conditional on ongoing monitoring of error rates, equity of access across patient groups, and impact on downstream demand. If monitoring shows rising safety incidents or systematic bias as the system is updated or used more widely, predefined stop criteria would trigger review, restriction or withdrawal. This allows

NICE guidance to remain responsive to real-world evidence, rather than relying solely on assumptions made at the point of appraisal.

24. Relative feasibility and practicality of the recommendations for NICE implementation

Although the recommendations set out above are intended to operate as a coherent package, they differ substantially in how feasible and practical they are for NICE to implement within existing HTA methods, processes and institutional constraints. Some recommendations largely formalise or extend practices that NICE already applies in other appraisal contexts, while others would require new forms of evidence, new relationships with developers and deployers, or enhanced post-recommendation oversight. This section therefore ranks the recommendations according to, in the authors opinion, their relative ease of implementation for NICE focusing on what can be adopted most readily within current processes and where more substantive development would be required.

24.1. Tier 1: High feasibility within existing NICE processes

24.1.1. Lifecycle modelling and explicit lifecycle costing in economic appraisals

This recommendation is among the most immediately feasible for NICE to implement because it builds directly on established health economic principles rather than requiring new evaluative frameworks. NICE already expects economic models to reflect ongoing costs, uncertainty over time, and realistic patterns of use where these are decision relevant. Extending this expectation explicitly to frontier AI, by requiring lifecycle costing that incorporates monitoring, governance, update management, assurance activities, and changes in utilisation, represents a clarification and strengthening of existing practice rather than a methodological change.

Comparable approaches are already used by NICE when assessing technologies delivered as services rather than discrete products, including managed diagnostic services, screening programmes and digital monitoring platforms. In these cases, committees routinely consider recurrent costs, staff time, downstream resource use and sustainability over time. Frontier AI raises the same issues, but more consistently and at greater scale. As a result, this recommendation can be operationalised largely through updated guidance to submitters and committees, without requiring new institutional capability.

24.1.2. Stronger and more explicit comparator justification at scoping

Requiring explicit justification of comparator choice at scoping is also highly feasible, as it relies primarily on procedural clarity rather than new analytical tools. NICE already treats comparator selection as a central decision in appraisal, particularly where technologies alter pathways, redistribute tasks or change service organisation. Frontier AI accentuates this issue, but does not fundamentally alter it.

In many existing appraisals, such as diagnostics, pathway innovations and service-level interventions, NICE already requires committees to interrogate whether the proposed comparator reflects real NHS practice and whether claimed benefits arise from genuine incremental effects. Making this expectation explicit and systematic for frontier AI can therefore be achieved through strengthened scoping guidance and committee practice, rather than through structural change.

24.2. Tier 2: Moderate feasibility with targeted procedural development

24.2.1. Binding guidance to a specific version, configuration and scope of use, with predefined review triggers

This recommendation aligns well with NICE's existing use of conditional recommendations, defined indications and review points, but requires more systematic application for frontier AI. NICE already links guidance to specific populations, indications and modes of use, and in some cases specifies circumstances under which recommendations should be reviewed.

What is novel here is the need to anchor recommendations to configuration and lifecycle characteristics rather than solely to clinical indication. Implementing this does not require a wholesale change to NICE processes, but it does require clearer conventions for specifying what exactly has been appraised and for identifying changes that are sufficiently material to trigger review. The approach is therefore feasible, but would benefit from piloting and iterative refinement to ensure consistency and proportionality.

24.2.2. Structured transferability analysis and UK-relevant real-world evidence planning

NICE routinely considers the generalisability of evidence, particularly where data are derived from non-UK settings or highly controlled environments. Requiring structured transferability analysis for frontier AI builds directly on this practice, but is more demanding because performance, risk and costs are highly sensitive to local digital maturity, workflows and governance arrangements.

Implementing this recommendation would likely require clearer templates or expectations for submitters, to ensure that transferability is addressed consistently and meaningfully rather than superficially. However, it does not require new HTA methods and can be integrated into existing appraisal workflows with targeted guidance.

24.2.3. Expanded use of proportionate conditional recommendations and maintained guidance

NICE already operates a proportionate approach to uncertainty through established conditional and managed access routes, including evidence-generation requirements and time-limited recommendations in areas where the evidence base is still developing (for example, through managed access arrangements such as the Cancer Drugs Fund), and is also progressing a more explicit whole-lifecycle approach to guidance maintenance. Building on these existing mechanisms to cover frontier AI is therefore conceptually straightforward and consistent with current practice, including options to support earlier decision-making (for instance, through early value assessment approaches where appropriate) alongside clear requirements for post-market evidence generation. The additional challenge is the expected pace and character of change in frontier AI, such as rapid model updates, altering performance across settings, and evolving risks, which may necessitate more active monitoring, clearer update and version-control expectations, and more frequent reassessment than many current managed access arrangements typically assume. This makes the conditional recommendation feasible but operationally more complex, requiring careful design to avoid disproportionate administrative burden and to set out unambiguous responsibilities for monitoring, evidence generation, data access, and the triggers for review or withdrawal of recommendations.

24.3. Tier 3: Lower feasibility and longer-term development

24.3.1. Explicit treatment of procurement, cybersecurity and assurance burdens as cost- and feasibility-relevant

Although NICE already considers feasibility and implementation issues in appraisal, making procurement complexity, cybersecurity requirements and assurance burdens explicit and systematic components of cost and feasibility assessment for frontier AI is more challenging than it may initially appear. Unlike many established technologies, these burdens are not fixed or uniform, but vary substantially across NHS organisations depending on digital maturity, local governance arrangements, procurement routes, and existing contractual relationships with suppliers and platform providers.

For frontier AI in particular, procurement and cybersecurity requirements may evolve over time as systems are updated, integrated with additional data sources, or deployed across new settings. Assurance activities such as penetration testing, model monitoring, audit, incident management and supplier oversight may therefore represent ongoing, variable and partially unpredictable costs rather than one-off implementation considerations. Capturing these burdens in a way that is both proportionate and decision relevant would require more than simple cost enumeration, and may necessitate new conventions for describing and modelling organisational capability, risk exposure and ongoing governance effort.

In practice, this recommendation is therefore not entirely straightforward to implement. While it does not introduce wholly new evaluative domains, it does require more explicit expectations about how submitters characterise procurement models, cybersecurity dependencies and assurance responsibilities, and about how committees interpret and compare these across technologies and settings. It may also require closer alignment between NICE appraisal processes and NHS procurement, digital and cybersecurity functions, to ensure that assumptions about feasibility and cost reflect operational reality rather than idealised deployment scenarios.

24.3.2. Structured frontier AI system specification as a condition of appraisal

Although foundational to many of the other recommendations, this is more challenging to implement because it depends on standardising information that is not yet consistently available. Unlike economic modelling or comparator justification, system specification requires agreement on how to describe configuration, workflow integration, dependencies and governance arrangements in a way that is both meaningful and proportionate.

NICE already requires detailed technology descriptions, but extending this to a structured, frontier-AI-specific specification represents a substantive change to appraisal requirements. It is feasible, but likely to require phased introduction, iteration and alignment with regulatory and international transparency initiatives, rather than immediate full implementation.

24.3.3. Predetermined change control plans or equivalent HTA change plans

This recommendation is conceptually aligned with emerging regulatory practice for adaptive AI, but presents practical challenges for NICE because it implies an ongoing relationship between appraisal, post-deployment change and review. Unlike static recommendation conditions, change plans require mechanisms to track updates, assess their significance and determine whether re-appraisal is needed.

While NICE has experience with evidence-generation requirements and conditional access, embedding structured change control into HTA would require new processes and potentially new data flows. This places it firmly in the category of medium- to longer-term development rather than near-term implementation.

24.3.4. Building and sustaining multidisciplinary assessment capability

Developing consistent multidisciplinary capability for frontier AI assessment is inherently resource-intensive and therefore among the least immediately implementable recommendations. Although NICE already draws on a wide range of expertise, frontier AI places ongoing demands on technical, organisational, behavioural and governance understanding that go beyond current norms.

This recommendation is best understood as an enabling condition for the others rather than a discrete procedural change. It is critical for long-term robustness and consistency, but will require investment, learning and institutional development over time rather than rapid implementation.

25. Recommendations linked to NICE process stages

The table below shows how each recommendation can be put into practice at different stages of the NICE process: scoping, committee review, managed access or conditional recommendation, and surveillance or maintained guidance.

Table 1: Recommendations linked to practical actions

Recommendation	Scoping	Committee review	Conditional recommendation	Surveillance / maintained guidance
Specify precisely what the intervention is to be evaluated	Require a structured system description that defines what is being assessed, including intended use, users, pathway role, setting and exclusions.	Use the specification to test boundaries, dependencies and assumptions; base conclusions on the system-in-use.	Use the specification as the baseline for deployment conditions and evidence-generation requirements.	Use the specification to detect configuration drift, scope creep or unevaluated repurposing.
Bind recommendations to version, configuration and scope, with defined review triggers	Define the evaluated version, configuration and scope; agree which changes could affect decisions.	State clearly that guidance applies only to the evaluated configuration; agree what counts as a material change.	Make continued access conditional on staying within scope and notifying changes.	Maintain guidance using pre-specified triggers linked to updates, drift, incidents or scope changes.
Require PCCP-like change governance plans or equivalent HTA change plans	Request an overview of how changes will be managed and what types of modification are anticipated.	Review whether clear plans are in place for how important updates will be checked, recorded and reported, and whether these give the committee enough clarity on when a recommendation may need to be revisited	Make change controls, notifications, audits and stop rules conditions of coverage for high-impact systems.	Monitor compliance with the change plan and trigger review when thresholds are crossed.
Strengthen transparency requirements	Specify required transparency materials (purpose, limits, subgroup gaps, uncertainty) and require consistency across artefacts.	Use these materials to assess limitations, failure modes, usability and human-factors risks.	Make user-facing disclosures, update notices and safety communications conditions of access.	Use transparency outputs to support surveillance, incident learning and guidance updates.
Add a short frontier AI checklist to NICE guidance to ensure consistent reporting.	Apply the annex as a checklist to ensure all frontier-AI-relevant domains are in scope.	Use a simple checklist to distinguish what NICE should assess, what should be covered by existing approvals, and what should be reflected in conditions for us.	Link access to targeted evidence generation where annex domains are weak or incomplete.	Use annex domains as standard surveillance items (e.g. bias drift, cyber events, training compliance).
Strengthen NICE's capacity to assess frontier AI and standardise multidisciplinary assessment processes.	Trigger frontier-AI pathways and assign internal methods support early.	Support committees with standard templates, consistent challenge questions and access to expertise.	Provide central support for monitoring and evidence governance during managed access.	Maintain shared learning, update taxonomies and consistent surveillance across appraisals.
Evaluate the human–AI system and sociotechnical fit as a core appraisal object	Specify users, workflow integration, autonomy level, oversight design and competency requirements.	Review evidence on human–AI interaction, escalation pathways and mitigation of automation bias.	Require training, local oversight arrangements and monitoring of real-world use patterns.	Monitor sociotechnical risks such as misuse, over-reliance, workflow disruption or documentation errors.

Recommendation	Scoping	Committee review	Conditional recommendation	Surveillance / maintained guidance
Require transferability analysis and UK-relevant real-world evidence planning	Define NHS pathways, settings and populations; specify key transferability questions and data needs.	Assess whether evidence generalises to NHS practice; scrutinise post-deployment evidence plans.	Use the plan to guide real-world evidence collection during managed access.	Use UK real-world data to update guidance or trigger reappraisal where transferability fails.
Strengthen comparator expectations and enforce explicit comparator justification at scoping	Lock comparators to current NHS practice and service models; define the counterfactual pathway.	Test whether comparator choice is appropriate and whether outcomes map to downstream effects.	Condition access on comparator-relevant evidence where uncertainty remains.	Monitor whether real-world practice continues to match the assumed comparator.
Take account of the real work and costs of putting frontier AI into practice.	Identify procurement routes, cyber requirements, infrastructure needs and site readiness constraints.	Assess feasibility, interoperability, assurance duplication and operational burden.	Make rollout conditional on meeting cyber and infrastructure prerequisites.	Monitor cyber incidents, interoperability failures and assurance workload over time.
Strengthen economic evidence to reflect lifecycle use and uncertainty for high-update frontier AI.	Specify the economic questions, time horizon, lifecycle costs and productivity mechanisms.	Test assumptions about uptake, updates, governance intensity and uncertainty.	Link access to budget-impact monitoring, utilisation limits or evidence-based pricing where needed.	Use real-world costs, use patterns and outcomes to update models and reassess value.
Use proportionate conditional recommendation pathways and maintained guidance for high-update frontier AI	Identify candidates for conditional routes based on update frequency, risk and evidence maturity.	Agree conditions, evidence milestones, stop criteria and review timelines.	Operate managed access with clear continuation, suspension or withdrawal rules.	Maintain guidance with scheduled and event-triggered reviews linked to updates and incidents.

26. Conclusion

Frontier AI does not require NICE to abandon its core principles of evidence-based decision-making, value for money, and equity. Instead, it requires those principles to be applied through a lifecycle lens that recognises rapid technical change, sociotechnical complexity, and ongoing governance as inherent features of this class of technology, rather than as downstream implementation concerns.

The central challenge is that frontier AI unsettles several assumptions on which conventional appraisal depends. Systems may change after evaluation, leading to evidence that becomes less relevant over time. Weak specification and limited transparency can undermine reproducibility and comparability. Comparator choice and evidence transferability are often complex, because frontier AI may reshape care pathways rather than substitute neatly for existing practices. Strong technical performance does not reliably translate into real-world benefit without careful attention to workflow integration, professional behaviour, and organisational context. Safety and effectiveness frequently depend on how humans and AI systems work together, while ethics, equity, cybersecurity, organisational readiness, and operational constraints become parts of the intervention cost rather than background context.

Frontier AI also presents distinctive challenges for health economic evaluation. Prices and patterns of use may change over time rather than remain fixed. Learning effects and organisational adaptation influence when and how value is realised. Governance, monitoring, and assurance generate ongoing costs rather than one-off expenditures. System-wide effects, such as induced demand or pathway feedback, can alter costs and benefits across services. Uncertainty and heterogeneity may be structural, arising from the evolving nature of the technology itself (for example as scope, performance, or deployment models change) rather than uncertainty that can be resolved simply by collecting more data in trials. Taken together, these features mean that static, point-in-time appraisal risks misestimating both value and cost, and under-characterising the types and distribution of risk introduced by frontier AI.

The recommendations in this report respond to these challenges in a pragmatic and operational way by focusing on how frontier AI can be made evaluable and governable in practice. Central to this approach are explicit system specification, recommendations that are clearly bound to version, configuration and scope, and structured approaches to managing change over time. Transparency documentation, such as clear descriptions of system function, limitations, update processes, performance variation, and post-deployment monitoring, should be treated as relevant evidence rather than optional background material. Assessment could be supported by a frontier-AI-specific annex with minimum reporting expectations, alongside stronger requirements around comparator choice, transferability to NHS settings, and sociotechnical fit. Procurement constraints, cybersecurity, interoperability and assurance burdens should be addressed explicitly as determinants of feasibility and value for money, and economic evaluation could reflect lifecycle costs, dynamic use and persistent uncertainty. Together, these considerations support the use of proportionate conditional recommendation pathways and maintained guidance for high-update systems, as well as sustained investment in NICE’s multidisciplinary capability.

By clearly defining what is being appraised, anchoring recommendations to how frontier AI systems are actually used and how they evolve over time, and embedding lifecycle thinking within HTA processes, NICE can continue to enable valuable innovation while protecting patients, maintaining public trust, and supporting the long-term sustainability of the health and care system.

27. REFERENCES

1. Association of British HealthTech Industries (ABHI). From policy to practice: using digital health technologies to address current NHS challenges. London: ABHI; 2023. Available from: https://www.abhi.org.uk

2. Association of British HealthTech Industries (ABHI). Perspectives on delivery: Life Sciences Sector Plan and 10 Year Health Plan. London: ABHI; 2025. Available from: https://www.abhi.org.uk

3. Ahmed M, Dai T, Channa R, Abramoff MD, Lehmann HP, Wolf RM. Cost-effectiveness of AI for pediatric diabetic eye exams from a health system perspective. npj Digit Med. 2025;8:3. Available from: https://doi.org/10.1038/s41746-024-01382-4

4. Alami H, Lehoux P, Denis JL, Motulsky A, Petitgand C, Savoldelli M, et al. Organizational readiness for artificial intelligence in healthcare: insights for decision-making and practice. J Health Organ Manag. 2020;34(1):106–14. Available from: https://doi.org/10.1108/JHOM-09-2019-0240

5. Aquino YSJ, Rogers WA, Jacobson SLS, Richards B, Houssami N, Woode ME, et al. Defining change: exploring expert views about the regulatory challenges in adaptive artificial intelligence for healthcare. Health Policy Technol. 2024;13(3):100892. Available from: https://doi.org/10.1016/j.hlpt.2024.100892

6. AXREM (Association of Healthcare Technology Providers for Imaging, Radiotherapy and Care). AXREM AI Special Focus Group manifesto. London: AXREM; 2025. Available from: https://www.axrem.org.uk

7. AXREM (Association of Healthcare Technology Providers for Imaging, Radiotherapy and Care). AXREM State of the Nation report. London: AXREM; 2025. Available from: https://www.axrem.org.uk

8. Azad TD, et al. Lessons from Henrietta Lacks inform a transparency framework to catalyze generative artificial intelligence in medicine. npj Digit Med. 2025;8:280. Available from: https://doi.org/10.1038/s41746-025-01526-4

9. Bélisle-Pipon JC, Couture V, Roy MC, Ganache I, Goetghebeur M, Cohen IG. What makes artificial intelligence exceptional in health technology assessment? Front Artif Intell. 2021;4:736697. Available from: https://doi.org/10.3389/frai.2021.736697

10. Bommasani R, Singer SR, Appel RE, Cen S, Cooper AF, Cryst E, et al. The California Report on Frontier AI Policy. Sacramento: State of California; 2025.

Available from: https://www.gov.ca.gov/wp-content/uploads/2025/06/June-17-2025-%E2%80%93-The-California-Report-on-Frontier-AI-Policy.pdf

11. Boverhof BJ, Redekop WK, Visser JJ, Uyl-de Groot CA, Rutten-van Mölken MPMH. Broadening the HTA of medical AI: a review of the literature to inform a tailored approach. Health Policy Technol. 2024;13(2):100868. Available from: https://doi.org/10.1016/j.hlpt.2024.100868

12. Bujkiewicz, S., Achana, F., Papanikos, T., Riley, R.D., & Abrams, K.R. 2019. NICE DSU Technical Support Document 20: Multivariate meta-analysis of summary data for combining treatment effects on correlated outcomes and evaluating surrogate endpoints.

13. Canadian Agency for Drugs and Technologies in Health (CADTH). An overview of continuous learning artificial intelligence-enabled medical devices. CADTH Horizon Scan. Ottawa: CADTH; 2022.

14. Canada's Drug Agency. 2025 watch list: artificial intelligence in health care. Ottawa: CDA-AMC; 2025. Available from: https://www.cda-amc.ca

15. Department for Science, Innovation and Technology (UK). Frontier AI: capabilities and risks – discussion paper. London: UK Government; 2023. Available from: https://www.gov.uk

16. Di Bidino R, Daugbjerg S, Papavero SC, Haraldsen IH, Cicchetti A, Sacchini D. Health technology assessment framework for artificial intelligence-based technologies. Int J Technol Assess Health Care. 2024;40:e61. Available from: https://doi.org/10.1017/S0266462324000308

17. Dias, S., Welton, N. J., Sutton, A. J., & Ades, A. E. (2011; last updated 2016). NICE DSU Technical Support Document 2: A generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials. NICE Decision Support Unit / University of Sheffield. https://sheffield.ac.uk/media/34176/download?attachment=

18. Duggan MJ, et al. Clinician experiences with ambient scribe technology to assist with documentation burden and efficiency. JAMA Netw Open. 2025. Available from: https://jamanetwork.com

19. Elvidge J, Hawksworth C, Avşar TS, Zemplenyi A, Chalkidou A, Petrou S, et al. CHEERS-AI: consolidated health economic evaluation reporting standards for interventions that use artificial intelligence. Value Health. 2024;27(9):1196–205. Available from: https://doi.org/10.1016/j.jval.2024.05.006

20. Farah L, Borget I, Martelli N, Vallée A. Suitability of the current health technology assessment of innovative artificial intelligence-based medical devices: a scoping review. J Med Internet Res. 2024.

21. Faria, R., Hernandez Alava, M., Manca, A., & Wailoo, A. J. (2015). NICE DSU Technical Support Document 17: The use of observational data to inform estimates of treatment effectiveness in technology appraisal: Methods for comparative individual patient data. NICE Decision Support Unit / University of Sheffield. https://sheffield.ac.uk/sites/default/files/2022-02/TSD17-DSU-Observational-data-FINAL.pdf

22. Fasterholdt I, Naghavi-Behzad M, Rasmussen BSB, Kjølhede T, Skjøth MM, Hildebrandt MG, Kidholm K. Value assessment of artificial intelligence in medical imaging: a scoping review. BMC Med Imaging. 2022;22:187. Available from: https://doi.org/10.1186/s12880-022-00918-y

23. Food and Drug Administration (US). Marketing appraisal recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions. Silver Spring (MD): FDA; 2025. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-appraisal-recommendations-predetermined-change-control-plan-artificial-intelligence

24. Ghabri S. Using AI in the economic evaluation of AI-based health technologies. PharmacoEconomics. 2025;43(6):597–600. Available from: https://doi.org/10.1007/s40273-025-01496-x

25. Haberle T, Cleveland C, Snow GL, et al. Impact of Nuance DAX ambient listening AI documentation: a cohort study. J Am Med Inform Assoc. 2024;31(4):975–9. Available from: https://doi.org/10.1093/jamia/ocae022

26. Hayat H, Kudrautsau M, Makarov E, et al. Toward the autonomous AI doctor. medRxiv. 2025. Available from: https://doi.org/10.1101/2025.07.14.25331406

27. Hu D, et al. Generative AI for drafting responses to patient messages: a systematic review of early evidence. npj Digit Med. 2025. Available from: https://doi.org/10.1038/s44401-025-00032-5

28. International Network of Agencies for Health Technology Assessment (INAHTA). INAHTA position statement: disruptive technologies. Edmonton: INAHTA; 2022. Available from: https://www.inahta.org

29. Jacob C, Brasier N, Laurenzi E, et al. AI for IMPACTS framework. J Med Internet Res. 2025;27:e67485. Available from: https://doi.org/10.2196/67485

30. Lanyi K, Twentyman K, Forsythe I, Green E, Barnabas J, Mkwashi A. Horizon scan of emerging generative AI-enabled technologies for healthcare. London: NIHR Innovation Observatory; 2025.

31. McDuff D, et al. Towards accurate differential diagnosis with large language models. Nature. 2025;642:451–7. Available from: https://doi.org/10.1038/s41586-025-08869-4

32. Medicines and Healthcare products Regulatory Agency (UK), Food and Drug Administration (US), Health Canada. Transparency for machine learning-enabled medical devices: guiding principles. 2024.

33. Moor M, Banerjee O, Shakeri Hossein Abad Z, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616:259–65. Available from: https://doi.org/10.1038/s41586-023-05881-4

34. National Institute for Health and Care Excellence (NICE). Evidence standards framework for digital health technologies. London: NICE; 2022. Available from: https://www.nice.org.uk/what-nice-does/digital-health/evidence-standards-framework-esf-for-digital-health-technologies

35. National Institute for Health and Care Excellence (NICE). (2022, 23 June). NICE real-world evidence framework (Corporate document ECD9). NICE.

36. National Institute for Health and Care Excellence (NICE). (2022, 23 June). NICE real-world evidence framework (Corporate document ECD9) : chapter: Methods for real-world studies of comparative effects. NICE. https://www.nice.org.uk/corporate/ecd9/chapter/methods-for-real-world-studies-of-comparative-effects

37. National Institute for Health and Care Excellence (NICE). (2025). NICE technology appraisal and highly specialised technologies guidance: the manual (PMG36). NICE. https://www.nice.org.uk/guidance/pmg36/resources/nice-technology-appraisal-and-highly-specialised-technologies-guidance-the-manual-pdf-72286779244741

38. Organisation for Economic Co-operation and Development (OECD). AI in health: huge potential, huge risks. Paris: OECD Publishing; 2024. Available from: https://www.oecd.org

39. Oettl FC, Zsidai B, Oeding JF, Samuelsson K. Artificial intelligence and musculoskeletal surgical applications. HSS J. 2025. Available from: https://doi.org/10.1177/15563316251339596

40. Oyewole A, Marshall C, Robertson E, et al. Identification of AI technologies in healthcare. London: NIHR Innovation Observatory; 2021.

41. Panda PK. Ethical use of AI in infectious diagnostic decision and therapeutic stewardship. IDCases. 2025;42:e02356.

42. Phillippo, D. M., Ades, A. E., Dias, S., Palmer, S., Abrams, K. R., & Welton, N. J. (2016). NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in appraisals to NICE. NICE Decision Support Unit / University of Sheffield. https://sheffield.ac.uk/media/34216/download

43. Singhal K, et al. Toward expert-level medical question answering with large language models (Med-PaLM). 2025.

44. van Kessel R, Schmidt J, Winitsky S, Wharton G, Mossialos E. Evaluation framework for health professionals' digital health and AI technologies. London: LSE Consulting; 2025. Available from: https://doi.org/10.21953/lse.vi7ayokh6s1u

45. Vielhauer J, Ruzicka M, Benesch C, et al. Development and validation of a high-stakes decision support agent in a clinical agentic AI system. SSRN Preprint. 2025. Available from: https://ssrn.com/abstract=5317661

46. World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: WHO; 2024. Available from: https://www.who.int

APPENDICES

Appendix 1: Detailed Search Strategy

A systematic search of relevant databases was conducted in November 2025, with the aim of retrieving published academic literature relating to the research question. The search strategy included keywords focusing on the following concepts: AI (including frontier AI); HTA, cost effectiveness and economic evaluations; medical technology. The search terms related to the AI concept were largely based on two published filters (Ayiku et al. in unmodified form; Campbell et al. in an adapted form to focus more on relevant technology for this question). Additional terms were added to specifically include the frontier or adaptive AI concept. No geographical, date or publication language limits were applied.

The following table summarises the database searches that were conducted and the number of results retrieved. Individual search strategies for each database are below.

Database	**Platform**	**Search date**	**No. of results**
MEDLINE (ALL)	Ovid	05/11/25	3,668
Embase	Ovid	05/11/25	2,837
EconLit	Ovid	05/11/25	32
Total number of results following removal of duplicates			**4,407**

Ovid MEDLINE(R) ALL <1946 to November 04, 2025>

1 algorithm*.ti,kf. 85740
2 (algorithm* adj2 (learn* or automate* or detect* or predict* or treatment* or therap* or radiolog* or AI or DL or data or dataset* or base* or classif*)).ab. 121216
3 artificial intelligen*.ti,ab,kf. 90754
4 AI.ti,kf. 25799
5 (machine adj2 learn*).ti,ab,kf. 171191
6 machinelearn*.ti,ab,kf. 33
7 (deep adj2 learn*).ti,ab,kf. 100530
8 deeplearn*.ti,ab,kf. 43
9 neural network*.ti,ab,kf. 141672
10 (convolutional adj1 network*).ti,ab,kf. 5342
11 automate*.ti. 54716
12 (automate* adj3 (system* or score* or software* or analysis* or analyse* or risk* or evaluat* or tool* or detect* or process*)).ab,kf. 50298
13 (vector machine* or svm*).ti,ab,kf. 41586
14 radiomic*.ti,ab,kf. 16475

15 ((supervised or unsupervised) adj3 (classifier* or prediction*)).ti,ab,kf. 1164
16 (‘frontier AI’ or ‘frontier artificial intelligence’ or ‘adaptive AI’ or ‘adaptive artificial intelligence’).ti,ab,kf. 60
17 (‘generative pre-trained transformer’ or ‘generative pretrained transformer’ or gpt* or ‘large language model*’ or llm* or ‘natural language process*’ or ‘foundation model*’).ti,ab,kf. 31134
18 (agent* adj2 (AI or ‘artificial intelligence’)).ti,ab,kf. 433
19 (Perplexity or Runway AI or Runway Gen-1 or ‘Bing chat’ or ChatGPT* or ‘Chat GPT’ or ‘Google* Bard’ or ‘Google* Gemini’ or ‘IBM Watson’ or ‘Microsoft* Bing’ or ‘Microsoft* Copilot’ or OpenAI or ‘Open AI’ or PathAI or ‘Path AI’ or DeepSeek or Grok).ti,ab,kf. 9435
20 or/1-19 626917
21 Technology Assessment, Biomedical/ 11769
22 (hta or ‘health technology assessment*’).ti,ab,kf. 10300
23 (cost* adj2 (effective* or utilit*)).ab,kf. 218267
24 ((economic adj2 evaluation*) or safety or efficacy).ti,ab. 1847318
25 21 or 22 or 23 or 24 2037626
26 ‘Equipment and Supplies’/ 24797
27 ((medical adj2 (tech* or device*)) or healthtech or medtech or healthcare).ti,ab. 503872
28 26 or 27 525153
29 20 and 25 and 28 3668

Econlit <1886 to October 30, 2025>

1 algorithm*.ti,ab. 21764
2 (algorithm* adj2 (learn* or automate* or detect* or predict* or treatment* or therap* or radiolog* or AI or DL or data or dataset* or base* or classif*)).ab. 3195
3 artificial intelligen*.ti,ab. 2566
4 AI.ti. 788
5 (machine adj2 learn*).ti,ab.4752
6 machinelearn*.ti,ab. 0
7 (deep adj2 learn*).ti,ab. 809
8 deeplearn*.ti,ab. 0
9 neural network*.ti,ab. 3245
10 (convolutional adj1 network*).ti,ab. 19
11 automate*.ti. 662
12 (automate* adj3 (system* or score* or software* or analysis* or analyse* or risk* or evaluat* or tool* or detect* or process*)).ab. 561
13 (vector machine* or svm*).ti,ab. 618
14 radiomic*.ti,ab. 1
15 ((supervised or unsupervised) adj3 (classifier* or prediction*)).ti,ab. 15
16 (‘frontier AI’ or ‘frontier artificial intelligence’ or ‘adaptive AI’ or ‘adaptive artificial intelligence’).ti,ab. 2
17 (‘generative pre-trained transformer’ or ‘generative pretrained transformer’ or gpt* or ‘large language model*’ or llm* or ‘natural language process*’ or ‘foundation model*’).ti,ab. 715
18 (agent* adj2 (AI or ‘artificial intelligence’)).ti,ab. 20

19 (Perplexity or Runway AI or Runway Gen-1 or ‘Bing chat’ or ChatGPT* or ‘Chat GPT’ or ‘Google* Bard’ or ‘Google* Gemini’ or ‘IBM Watson’ or ‘Microsoft* Bing’ or ‘Microsoft* Copilot’ or OpenAI or ‘Open AI’ or PathAI or ‘Path AI’ or DeepSeek or Grok).ti,ab. 125

20 or/1-19 31520

21 (hta or ‘health technology assessment*’).ti,ab,kw. 252

22 (cost* adj2 (effective* or utilit*)).ab,kw. 7173

23 ((economic adj2 evaluation*) or safety or efficacy).ti,ab. 21644

24 21 or 22 or 23 28278

25 ((medical adj2 (tech* or device*)) or healthtech or medtech or healthcare).ti,ab. 7322

26 20 and 24 and 25 32

Embase <1974 to 2025 Week 44>

1 algorithm*.ti,kf. 105443

2 (algorithm* adj2 (learn* or automate* or detect* or predict* or treatment* or therap* or radiolog* or AI or DL or data or dataset* or base* or classif*)).ab. 158884

3 artificial intelligen*.ti,ab,kf. 105152

4 AI.ti,kf. 30209

5 (machine adj2 learn*).ti,ab,kf. 196886

6 machinelearn*.ti,ab,kf. 380

7 (deep adj2 learn*).ti,ab,kf. 114137

8 deeplearn*.ti,ab,kf. 216

9 neural network*.ti,ab,kf. 164919

10 (convolutional adj1 network*).ti,ab,kf. 5938

11 automate*.ti. 71888

12 (automate* adj3 (system* or score* or software* or analysis* or analyse* or risk* or evaluat* or tool* or detect* or process*)).ab,kf. 75273

13 (vector machine* or svm*).ti,ab,kf. 49503

14 radiomic*.ti,ab,kf. 22075

15 ((supervised or unsupervised) adj3 (classifier* or prediction*)).ti,ab,kf. 1448

16 (‘frontier AI’ or ‘frontier artificial intelligence’ or ‘adaptive AI’ or ‘adaptive artificial intelligence’).ti,ab,kf. 71

17 (‘generative pre-trained transformer’ or ‘generative pretrained transformer’ or gpt* or ‘large language model*’ or llm* or ‘natural language process*’ or ‘foundation model*’).ti,ab,kf. 37417

18 (agent* adj2 (AI or ‘artificial intelligence’)).ti,ab,kf. 489

19 (Perplexity or Runway AI or Runway Gen-1 or ‘Bing chat’ or ChatGPT* or ‘Chat GPT’ or ‘Google* Bard’ or ‘Google* Gemini’ or ‘IBM Watson’ or ‘Microsoft* Bing’ or ‘Microsoft* Copilot’ or OpenAI or ‘Open AI’ or PathAI or ‘Path AI’ or DeepSeek or Grok).ti,ab,kf. 10594

20 or/1-19 769409

21 exp biomedical technology assessment/ 20004

22 (hta or ‘health technology assessment*’).ti,ab,kf. 18154

23 (cost* adj2 (effective* or utilit*)).ab,kf. 300591

24 ((economic adj2 evaluation*) or safety or efficacy).ti,ab. 2863645

25 21 or 22 or 23 or 24 3123866

26 medical device/ 56544

27 ((medical adj2 (tech* or device*)) or healthtech or medtech or healthcare).ti,ab. 762826
28 26 or 27 808011
29 20 and 25 and 28 5664
30 limit 29 to (embase or ‘preprints (unpublished, non-peer reviewed)’) 2837

The results of the database searches were initially screened at title and abstract level to assess relevance to frontier artificial intelligence specifically.

Grey literature searching was conducted in October and November 2025 and targeted a range of website sources with three main objectives. The first objective was to identify existing guidance from international health technology assessment organisations on the evaluation of AI-based medical technologies. The second was to identify examples of health technology assessments conducted on frontier AI medical technologies, or on AI-based medical technologies of potential relevance to frontier AI. The third objective was to undertake horizon scanning with a business and commercial focus, in order to identify frontier AI technologies likely to be considered in future health technology assessments or to emerge more broadly within the medical field over the coming years.

Websites were searched using keywords related to artificial intelligence, including ‘artificial intelligence’, ‘AI’, ‘machine learning’, and ‘large language models’. Where possible, filters were applied to restrict results to the period from 2023 to 2025, reflecting the rapid pace of development and the relative novelty of frontier AI technologies. The following table lists the website sources searched as part of the grey literature identification process.

Website	URL
HTA organisations and databases	
Agency for Healthcare Research and Quality (AHRQ)	https://www.ahrq.gov/
Canada’s Drug Agency (CDA-AMC)	https://www.cda-amc.ca/
European Commission (Health technology assessment)	https://health.ec.europa.eu/health-technology-assessment_en
European Medicines Agency (EMA)	https://www.ema.europa.eu/en/homepage
HTA Database (via INAHTA)	https://database.inahta.org/
International Network of Agencies for Health Technology Assessment (INAHTA)	https://www.inahta.org/
Institute for Quality and Efficiency in Health Care (IQWiG)	https://www.iqwig.de/en/

ISPOR	https://www.ispor.org/home
Medicines & Healthcare products Regulatory Agency (MHRA)	https://www.gov.uk/government/organisations/medicines-and-healthcare-products-regulatory-agency
NHS EED (via CRD)	https://www.crd.york.ac.uk/CRDWeb/
NICE	https://www.nice.org.uk/
U.S. Food & Drug Administration (FDA)	https://www.fda.gov/medical-devices
World Health Organization (WHO)	https://www.who.int/
Sources for upcoming medtech	
AIthority	https://aithority.com/
Association of British HealthTech Industries (ABHI)	https://www.abhi.org.uk/
Axrem	https://www.axrem.org.uk/
Digital Health	https://www.digitalhealth.net/
Fierce Healthcare	https://www.fiercehealthcare.com/ai-and-machine-learning
Hopkins Business of Health Initiative (HBHI)	https://hbhi.jhu.edu/
NIHR Innovation Observatory (NIHRIO)	https://io.nihr.ac.uk/

Appendix 2: Detailed summary of studies

The appendix tables are intended to be read together. Table A provides a structured summary of each included study, including the document type, the AI technology archetype, how frontier artificial intelligence is defined or implied, the key evaluation challenges identified by the authors, proposed solutions where available, and a concrete illustrative example drawn directly from the study. Table B synthesises these findings across studies by mapping identified challenges to health technology assessment domains and classifying them as technical or procedural in nature. Table B also links each challenge type to the types of evidence required for health technology assessment and to corresponding process actions, such as monitoring, re-review cadence, or conditional approval mechanisms. Together, the tables are designed to illustrate how frontier AI characteristics translate into specific evaluation, governance, and implementation implications for health systems.

All 37 documents identified through the search strategy were reviewed in full. Due to time and resource constraints, only empirical studies and preprints identified through the initial database searches were summarised in detail in the appendix tables. Studies and documents retrieved through supplementary searches of the grey literature, including those requested by the NICE advisory group, were reviewed and considered in the synthesis but were not tabulated in the appendix.

In total, 20 studies, comprising 18 peer-reviewed studies or reviews and 2 preprints, are presented in the appendix summary tables. These studies report on specific artificial intelligence systems, applications, or evaluative approaches and were selected for tabulation because they provide concrete examples of frontier or adaptive AI deployment, assessment, or performance. The remaining 17 documents consist primarily of guidance, policy, regulatory, and methodological framework publications. These were identified through targeted grey literature searching and supplementary identification and were not summarised in the appendix tables, but instead informed the narrative synthesis, contextual analysis, and interpretation presented in the main report.

Data extraction for studies included in the appendix tables followed a structured and transparent approach aligned to the table format. For each study, a single key illustrative example was identified, presented as a direct quotation from the source article, that best demonstrated challenges associated with the evaluation, regulation, or real-world deployment of frontier or adaptive artificial intelligence systems. For each study, information was extracted on how frontier artificial intelligence was defined or implied within the study. Where studies did not explicitly use the term frontier AI, aspects such as adaptivity, continuous learning, autonomy, generative capability, interoperability, short product lifecycles, or post-deployment model change were used to characterise frontier-like behaviour, as reflected in the appendix tables.

Challenges identified within each study were extracted and organised thematically and structurally in the appendix tables. Challenges are categorised by health technology assessment domain and further classified as technical or procedural in nature, reflecting whether they arise primarily from model behaviour and performance characteristics or from processes of evaluation, governance, implementation, and review. Extracted challenges span evidence generation and validation, clinical effectiveness, safety and incident management, equity, transparency and bias, interoperability and data integration, governance and re-review cadence, and economic evaluation.

Where studies proposed specific solutions or mitigation strategies to address identified challenges, these were extracted directly and summarised in the appendix tables. Where challenges were described without accompanying solutions, proposed actions were developed by the review team for each challenge to complete the table.

Study: Suitability of the Current Health Technology Assessment of Innovative AI-Based Medical Devices: Scoping Literature Review (J Med Internet Res, 2024).

Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
Suitability of the Current HTA of Innovative AI-Based Medical Devices: Scoping Review; 2024 (JMIR)	Scoping review (78 sources across 9 HTA domains)	No Frontier AI term. Findings discuss AI-MDs broadly. Possible frontier AI elements: AI-MDs have **short product lifecycles (around 12–18 months); some evolutive models (**evolutive deep learning–based medical device)**,** **data/**interoperability imply frontier-like behaviors ‘ability of AI-based medical devices to interact and communicate with other health care technologies and systems. This includes the ability to access and use data from different sources,	Diagnostic/triage imaging assistants; monitoring/ insulin delivery; pathway triage (e.g., screening ‘diabetic retinopathy screening AI-based medical device’)	Safety evidence gap, Transparency/ complexity/ interoperability (explainability), bias/fairness, data quality/representativeness (generalization), integration into workflows, and heterogeneous legal/ethical requirements.	Lifecycle HTA, dynamic (re)certification, RWE validation (require **version-linked evidence, continuous monitoring)** **Applied CEA:** include as costs training/IT/maintenance, monitoring/update, system impacts (capacity constraints, downstream utilisation, time-to-diagnosis, referral effects). Link HTA conclusions to AI **version** and plan re-assessment.	As an example, consider an AI tool used for diagnosing diabetic retinopathy in a primary care setting, such as by a family physician or nurse. In theory, this could lead to shorter waiting times for patients. However, if the health care organization faces challenges such as a shortage of specialized staff (eg, ophthalmologists), insufficient organization of care pathways, and lack of specialized facilities for proper management and follow-up after diagnosis, the introduction of AI might adversely affect the quality of care and patient experience. In such a

		such as laboratory systems, imaging archives, and patient health records.’				scenario, the AI application might merely transfer the delay from primary to secondary care, failing to address the fundamental issue.

MD: Medical Devices; US: United States; FDA: Food and Drug Administration; RWE: Real-world evidence

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What’s different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Evidence generation (SLR/RWE/validation)	Technical	**Rapidly evolving models** make prior validation stale; broader data dependence raises transportability issues.	**Version ID & change log**; external multi-site validation; RWE plan.	Version-bound recommendations; scheduled re-tests (6–12 months) or on major change or after model updates (the paper notes ‘dynamic certifications’ as mechanisms to re-evaluate after substantive modifications).
Clinical effectiveness (broad impact on patient health)	Technical	Frontier tools may alter **workflows and outcomes** across sites/subgroups as they update leading to broad system-based impacts on patient health	Decision-impact measures; subgroup analyses; prospective external validation; alignment to specific guidelines on their application.	Time-limited approvals; trigger re-review on performance drift.

Safety & incidents	Technical	**Safety evidence gap** and broader applications means new failure modes; risk profile can change post-update.	Safety plan; Adverse event taxonomy; post-market safety KPIs; monitoring protocol.	Continuous safety monitoring; pause/rollback rules; re-review after changes.
Equity, transparency & bias	Technical	Scale and data breadth heighten **bias/fairness** and transparency demands.	Evaluate fairness/bias, explainability, accountability, consent/privacy, cybersecurity, liability, and regulatory compliance as part of the HTA evidence package	Equity KPIs. Bias tests/mitigation plan; explainability/transparency docs.
Interoperability & data integration	Procedural	Compared to narrow AI there is heavy data dependence, need for interoperability and data integration (imaging, genomics, wearables), and stringent privacy/security.	ensure interoperability, data-integration capability, privacy/security compliance, and documentation of explainability/transparency (data used, model methods, limitations).	Site acceptance testing; gated rollout; standards conformance; security controls.
Governance & re-review cadence	Procedural	**Short lifecycles (12–18 months)** require ‘living’ HTA and dynamic (re)certification.	current HTA cycles struggle with frequent updates; need for ongoing monitoring/re-assessment; declared schedule of updates; version/evidence ‘expiry’; change plan.	Scheduled refresh. Trigger-based re-review (upon update, drift, incident). Coordination across stakeholders for data access/governance and implementation in real pathways.

Economics (costs & CEA scope)	Tech & Procedural	Larger **governance footprint** and refresh burden vs non-frontier; system effects material.	Full cost map: training, IT, maintenance, monitoring; pathway and downstream system impacts (e.g. queue impacts.). Secure infrastructure premium. Costs of recurring audits/re-tests and re-validation; assessor time; training/change in management, governance and monitoring. Validation/RWE costs. productivity gains.	Maintain **living** economic model; scenario tests for recurring audits/re-tests/update changes. Establish mechanisms to re-assess economic value over time as models update. Explore payer reimbursement strategies aligned to demonstrated value.

KPI: Key performance indicators.

Study: Health technology assessment framework for artificial intelligence-based technologies (Int J Technology Assessment in Health Care, 2024).

Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
Health technology assess	Study (2-round Delphi;	OECD AI system definition; recognises adaptive vs non-adaptive AI; not a	Broad AI (diagnosis, risk prediction,	HTA Core Model insufficient: 14/20	AI-adapted framework, early dialogue, lifecycle use and later	The AI-Mind platform is set to introduce two new AI-based tools: the AI-Mind Connector, which identifies

ment framework for AI-based technologies; 2024 (IJTAHC)	46 experts). Output: 48/65 topics rated 'critical' for AI-HTA (≥70% agreement).	'frontier AI' label. Outputs emphasize topics needed when AI may evolve (**fast-updating) and low-transparency models**; aspects related to Frontier AI	monitoring), including tools for early dementia (AI-Mind project context).	additional AI topics deemed critical (e.g., AI accuracy, data bias, risk management, benefit–harm ethics). Frontier systems' have heightened elements of **short lifecycles/adaptivity, opacity, and data bias risks**	testing (on the AI mind platform). Applied CEA**:** include process costs, monitoring, flag uncertainty/heterogeneity.	dysfunctional brain networks through high-density electroencephalographic recordings, and the AI-Mind Predictor, which assesses dementia risk using data from the Connector. These data include advanced digital cognitive tests, genetic and protein biomarkers, as well as important textual variables…. It is important to note that the assessment process will begin before the complete development of the AI-based tools. This requires …considering both early and comprehensive assessments. This nuanced approach aligns with the changing nature of AI development and emphasizes the need for adaptable HTA methodologies to suit evolving technological landscapes.

MD: Medical Devices; US: United States; FDA: Food and Drug Administration; RWE: Real-world evidence

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Health problem & current use	Procedural	Frontier AI can span multiple pathway roles and alter role over time (e.g., expand in scope from triage to autonomous management of individual patient pathways)	Explicit pathway map and boundaries; intended autonomy level.	Re-scope if model updates expand function or autonomy. In HTA costs include ddisplacement and de-implementation costs
Description & technical characteristics	Technical	Continuous-learning / multimodal models with less transparent architecture.	Versioned model include data sources, update schedule	Mandatory re-validation after each major model update. Include in HTA models vvalidation/Quality Assurance/ update costs.
Safety	Technical	Frontier AI has higher risk of silent failure and emergent behavior at scale.	Documented known failure modes, results of stress-testing under worst-case conditions, and evidence of how often humans have had to override the AI.	Real-time incident monitoring with rollback triggers. Include in HTA models monitoring program; expected incident loss.
Clinical effectiveness	Technical	Real-world performance may silently degrade over time or in different settings, without any automatic alert that the AI needs to be rechecked or retrained.	External validation, subgroup drift analysis, prospective pilot data.	Require adaptive re-assessment tied to real-world drift signals. Include in models FP/FN externalities.

Costs & economic evaluation	Technical	Frontier AI needs continuous retraining and infrastructure/compute, not one-off capital investment.	Full lifecycle costing (compute, retraining, human supervision, assurance).	Dynamic CEA with scheduled economic recalibration. Implementation, maintenance, training, process costs. Update Value of Information
Ethical analysis	Procedural	Opacity and autonomy create new risks for agency, consent, manipulation.	Provide explanations users can understand for the AI's outputs, a documented policy showing when and how humans can overrule the AI, and proof that patients/users are informed and have consent protections in place.	Ethics audit and governance checkpoint pre-release and post-update. Include in HTA models mitigation/compliance expenditure.
Organisational aspects	Procedural	Frontier AI may redistribute cognitive labour and create new dependencies on vendor.	Workflow modelling and staff skill impact and vendor lock-in analysis.	Implementation phased with organisational readiness gates. Include in HTA models capacity release and management change costs.
Patients & social	Procedural	Conversational/agentic models affect trust, identity, autonomy, and digital exclusion.	Acceptability studies including vulnerable groups; transparency that AI is being used for its intended purpose	Post-deployment experience monitoring and user feedback loops. Where possible include in models experience/access benefits (e.g. travel/wait-time benefits)

Legal aspects	Procedural	Frontier AI raises liability ambiguity and higher privacy/IP/security risks globally.	A documented data-protection impact assessment, a clear map of who is legally responsible for what (developer, deployer, clinician, organisation), and evidence that the system respects the laws/regulations of every place it will be used (with controls to prevent non-compliant use).	Legal/security audits and auto-triggered re-review on capability shift. Include ccompliance & insurance in HTA models (e.g. include in Budget Impact Analysis).

KPI: Key performance indicators.

Study: Marketing Appraisal Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions. FDA. August 18, 2025.

Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
Marketing Appraisal Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions. FDA. August 18, 2025.	Guidance for Industry and Food and Drug Administration Staff.	No explicit 'frontier AI.' The governance applies directly to rapidly updating AI.	AI-enabled devices that update continuously or manually; diagnostic monitoring aids; could include LLM.	Frontier systems update more often, scope creep in AI functionality	Bind decisions to Predetermined Change Control Plan version and acceptance criteria; require transparency & acceptance criteria. Need **version-linked appraisal, shorter re-review cycles,** and **public change logs/labels. Applied CEA:** recurrent costs for validation/monitoring/labeling/audits.	Optical Imaging System Co-packaged with Imaging Drug: The product is a device-led combination product including an AI based device integrated into an imaging system co-packaged with an approved optical imaging drug. The AI device analyses images in real-time and highlights potential cancerous lesions for further evaluation. The product was authorized with a Predetermined Change Control Plan.

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Update governance & versioning	Procedural	Frequent model changes may cause results to lose relevance; PCCP formalises allowed updates.	PCCP (title/version). Description of Modifications; Modification Protocol; Impact Assessment; version timeline/change log.	Bind guidance to authorized versions; re-review on out-of-scope change or acceptance-criteria fail.
Evidence binding to version	Technical	Frequent model changes demand version-specific performance & data traceability.	Traceability from version (including details on train/tune/test data); multi-site independent tests; preset acceptance criteria.	Version-linked recommendations; fail-trigger re-review; declare evidence 'expiry.'; Value of Information analysis of refresh.
Scope control & failure modes	Technical	Risk of scope creep with general-purpose models; new failure modes after updates.	Tight Description of Modifications; Impact Assessment including subgroup/equity effects; risk mitigations.	Block changes outside PCCP; pause/rollback if benefit–risk shifts; Incident probabilities/impacts; mitigation costs.

Transparency & labeling	Procedural	Users need to know when the model changed & how performance moved.	Labelling that device has an authorized PCCP; public summaries of implemented changes & performance.	Make public change log & label updates a coverage condition; check at refresh.
Supplier & release mode risk	Procedural	Heavier reliance on third-party external infrastructure or services, leading to dependency on their reliability, pricing, safety constraints, or even geopolitical risk.	Bill of materials (models/tools); update notice terms; update Quality Management System and FDA 'Q-Appraisal' (A process where you informally engage with the agency before submitting a full application) history (if available).	Supplier risk assessment; contingency paths if AI product scope shift; time-limited approval. Vendor management; include contingency/downtime scenarios in HTA models.
Monitoring & re-review cadence	Tech + Procedural	Higher update regularity	Monitoring metrics (drift, usage, equity); PCCP refresh plan/cadence; living HTA with automated monitoring.	Scheduled refresh (e.g., 6–12 months) and trigger-based re-review (version change, audit fail, incident); include continuing monitoring costs in HTA models.

Regulatory precedent for managed access	Procedural	Higher update regularity : FDA guidance allows in-scope updates without new appraisals, useful analogue for time-limited HTA.	Cite authorised PCCP status in appraisal; alignment with PCCP acceptance criteria.	Time-limited recommendations tied to PCCP milestones; include re-assessment costs in HTA models.

PCCP: Predetermined Change Control Plan.

Study: Ghabri, S., 2025. Using AI in the Economic Evaluation of AI-Based Health Technologies. *PharmacoEconomics*, pp.1-4.

Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
Using AI in the Economic Evaluation of AI-Based Health Technologies; 2025 (PharmacoEconomics, editorial	Editorial (methods/policy guidance)	No ‘frontier AI’ term; spans **generative AI/LLMs.**Specifically mentions these systems have rapid version cycles and broad use.	Two archetypes: (1) **AI-based HTs** (interventions under appraisal) , (2) **LLM assistants inside HTA**	**Dynamic tech & comparator versions**, reporting of AI specific learning curves, Skills gaps among clinicians/HE staff, and evolving pricing make **evidence go stale** faster	Use **CHEERS-AI**, plan **dynamic CEA (R**efresh ICERs as versions change; include implementation/ governance), and **govern the use of LLMs** in SLR/modelling with strict	It is reasonable to consider the characteristics identified by Tarricone et al. concerning mobile health (mHealth) apps, with adjustments accounting for the fast and dynamic nature of AI. Accordingly, the main considerations should

			(literature revies/modelling copilot)	and expand **governance/Quality Assurance**	Quality Control; skill up assessors.	be the adaptative nature that characterizes intrinsically digital technologies given the rapid evolution of AI tools, the more pronounced learning curve (compared with that of traditional medical devices), including interactions with users, the gradual organizational impact on healthcare practices and settings, and the dynamic pricing process resulting from the evolving nature of technologies affecting their efficacy and utilization.

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Evidence generation & reporting	Procedural	Frontier systems evolve quickly	Historic CE reporting is weak on implementation/costs so **needs stricter standards (**transparent tech costs & implementation detail.)	Mandate CHEERS-AI for AI-based appraisals; reject if key items missing. Include full implementation costs; sensitivity to reporting assumptions.

Dynamic technology & comparators	Technical	**Versions/generations and comparators change** faster than with non-frontier tools.	Version ID/generation history; comparator versioning; documented change cadence.	**Living** evaluation: scheduled CE refresh; tie recommendations to version. Schedule update scenarios; impact of comparator evolution on ICER.
Use of AI inside HTA methods (SLR/modelling)	Technical	LLMs assist search, extraction, modelling; risk of **errors/overfitting/poor generalisation.**	Human-in-the-loop plan (e.g. in Quality control); documentation of prompts/datasets; validation/replication artefacts.	Allow AI-assist only with Quality control (e.g. human oversight); audit trail required; escalate review for automation. Quality Assurance/review time; rework rates; delay costs vs productivity gains.
Data/model quality control	Technical	Frontier pipelines need **formal dataset & model Quality Control** (audits, validation) to avoid misuse.	Dataset documentation; audit schedule; validation addressing overfitting & generalisability.	Make Quality control artefacts a condition of acceptance, periodic audits. Quality control /audit costs; effect on uncertainty (wider/ narrower intervals).
Learning curves & organisational impact	Procedural	Strong **learning effects** and workflow change	Learning-curve assumptions; training/ throughput/time metrics.	Phased rollout with metrics; revisit model as productivity changes. Training & change-management costs; dynamic productivity gains.
Equity/affordability	Technical + Procedural	Risk of **inequitable access** and affordability	Equity relevant outcomes and sub	Monitor equity indicators; include affordability

		constraints magnified by rapid evolution.	groups; access constraints; pricing/affordability rationale.	checks in re-review. Distributional CEA; affordability scenarios/pricing dynamics.
Capability/skills in HTA teams	Procedural	Frontier use raises **skills gaps** in HTA, increasing misuse risk.	Staff training plans; named expertise for AI/Quality control oversight.	Require assessor upskilling; involve technical experts as needed. This involves training investment; reviewer time; error/misuse avoidance value.

PCCP: Predetermined Change Control Plan.

Study: Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles (UK government, June 2024)

Table A: summary of document

Source (short title; year)	Doc type	Frontier definition	Likely tech archetype	Summary of key Frontier challenges (how	Summary of key Frontier solutions	Example involving a case where AI technology faces

		used/ aspects of frontier AI	(e.g., diagnostic assistant)	HTA issues differ vs non-frontier)		a challenge relevant to frontier AI
Transparency for ML-enabled medical devices: guiding principles; 2024	UK government Guidance	No 'frontier AI' term; Applies to **frontier (LLM/LMM)** by requiring more frequent update notices and interface disclosures.	**LLM/LMM diagnostic/decision aids and LLM-based clinician copilots** embedded in devices	**Higher update tempo & general-purpose scope mean tighter, ongoing transparency** (change logs, on-screen guidance at high-risk steps, subgroup limits, confidence intervals) vs task-specific ML; supports **version-bound HTA and living re-review.**	**Tighter, ongoing transparency** (change logs, on-screen guidance at high-risk steps, subgroup limits)	No example given in the guidance

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Evidence communication & usability	Procedural	LLM/LMM tools need **audience-specific** info (clinician, patient, payer) and clearer risk/benefit logic than narrow ML.	**Transparency plan** mapping audiences; intended purpose; workflow role; inputs/outputs; effect on clinician judgement.	Make transparency documentation a **appraisal requirement;** verify in UI/training materials. Include in HTA models the costs of

				monitoring, and incident communications.
Performance & limitations	Technical	Broader scope means more **context-dependent limits**	Performance with uncertainty and (**confidence intervals)**; bias/limits; under-represented groups; contraindications.	Publish performance ranges; require equity sub group reporting; flag high-risk steps. Impact of uncertainty on ICER
Explainability / logic	Technical	Frontier models often **opaque**; still need usable ‘logic’ summaries for clinical decisions.	Available **logic/explainability** descriptions; basis of outputs; examples for critical decisions.	Accept concise logic narratives; tie to human-in-the-loop checkpoints.
Update notifications & change management	Procedural	**Frequent updates** require **timely notifications** and UI prompts at high-risk moments.	Update notification plan; **change log;** on-screen guidance placement.	Version-linked approvals; **shorter re-review cycles;** trigger-based alerts. Include in HTA models the costs of Notification/comms costs; re-validation & rollout overhead
Site acceptance & monitoring	Technical	Need **site-specific acceptance testing** and ongoing monitoring to sustain transparency over time.	Site validation guidance; monitoring metrics (drift, safety, equity); reporting cadence & responsibilities.	Mandate local acceptance tests; dashboarded monitoring; pause/rollback rules.
UI placement & human-centred design	Procedural	Frontier tools benefit from **in-product, just-in-time** instructions/alerts	Evidence of **UI-embedded** comms (tooltips/alerts), training	Require periodic interface review as part of re-assessment. Include in HTA models the costs of

			interface, and modality choices (text/video).	UI/content update costs; staff time to engage.

Study: AI for IMPACTS Framework for Evaluating the Long-Term Real-World Impacts of AI-Powered Clinician Tools: Systematic Review and Narrative Synthesis (J Med Internet Res, 2025).

Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
AI for IMPACTS framework for evaluating long-term real-world impacts of AI clinician tools; 2025 (JMIR)	Systematic review + narrative synthesis proposing **AI** framework (7 clusters; 28 subcriteria)	No ‘frontier AI’ term; adopts broad WHO view of AI in health care. **LLM/LMM/general-purpose tools** with rapid updates	Clinician-facing **decision support** (diagnosis/prognosis), **LLM report/summary assistants,** monitoring/triage tools	Greater need for **integration/workflow fit**, **ongoing monitoring/governance, LLM-specific performance checks,** and **explicit economic sustainability**	**L**ifecycle, monitoring, governance, assessor training	No case study given. Effective evaluation requires collaboration among professionals from various fields, such as medicine, IT, and social sciences to ensure a comprehensive assessment. This diversity of expertise is necessary to address the complexities of AI, from technical and ethical considerations to clinical relevance and real-world impact

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What’s different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Integration & workflow fit	Technical	General-purpose/LLM tools touch more workflow nodes and change faster than narrow ML means higher integration risk.	Evidence of interoperability, infrastructure needs, and workflow impact; site readiness/acceptance testing.	Gate deployments with site acceptance; declare evidence ‘expiry’ for integration claims; refresh after major updates. In HTA models cost Integration/IT costs; queue/throughput effects; training time.
Monitoring, governance & accountability	Procedural	Frontier pace amplifies drift/oversight burden	Monitoring plan (metrics, schedules), governance roles, update/maintenance procedures, accountability lines.	‘Living’ oversight with dashboards; shorter re-review cycles; documented roles/responsibilities. In HTA models cost ongoing monitoring/audit; rollback contingencies; staff time.
Performance & quality	Technical	Need LLM-specific checks (hallucination,) and checks	Foundational metrics and measures; external	Bind conclusions to version & tested prompts;

		of robustness/generalisation beyond accuracy.	multi-site validation; reliability/repeatability logs.	periodic re-tests under real-world data. In HTA models the cost of re-test costs.
Acceptability, trust & training	Procedural	Wider scope means higher cognitive load; clinicians need more training and just-in-time guidance.	User training/support plan; usability evidence; trust/acceptability measures.	Phased roll-out; UX reviews; refresh training post-update. In HTA models the cost of training & change-management; productivity trajectory (learning curve).
Cost & economic evaluation	Tech and Procedural	Governance and refresh schedule enlarge recurring costs vs non-frontier; economic sustainability must be shown.	full cost map including monitoring/maintenance; update cadence scenarios.	Maintain a 'living' economic model; scenario tests for drift/adoption. Include in HTA models the cost of Governance/audit; re-validation; long-run utilisation effects.
Technological safety & transparency	Technical	Broader use raise incident and explainability demands.	Safety plan; transparency of limits/failure modes; human-in-command controls.	Incident dashboards; pause/rollback triggers; verify explainability communications. Include in HTA models the cost of Incident probability/impact; downtime; mitigation programme costs.
Scalability & impact	Procedural	System-wide effects scale faster with frontier tools;	Evidence of clinical effectiveness/utility	Stage scale-up; monitor heterogeneous impacts;

		risk of uneven benefits across sites/populations.	across diverse settings/populations; scalability plan.	re-review on domain shift. Include in HTA models the cost of distributional impacts; capacity spillovers; environmental/compute costs (if relevant).

Study: Broadening the HTA of medical AI: A review of the literature to inform a tailored approach (Health Policy and Technology, 2024).

Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
Broadening the HTA of medical AI; 2024 (Health Policy & Technology)	Review and tailored assessment list (29 elements	No 'frontier' term. Paper targets medical AI with **continuous learning,**	Imaging diagnostic assistants ; **NLP/EHR abstraction** tools;	Short lifecycles & model drift, heavier **interoperability/ data** burden, and broader	Lifecycle HTA, **version-linked evidence, continuous monitoring, real-world validation**	No example case given in the paper. A medical AI technology may be affordable on a large scale but not in a smaller hospital ('economy of scale' principle).

	across 4 areas)	**interoperability, explainability, generalisability** gaps - features typical of frontier-leaning systems.	**ICU decision support** (three worked cases)	ethics/transparency needs	**across sites**, site acceptance, quasi-experimental designs and economic analysis that includes infrastructure/training/cyber/maintenance. Local Budget Impact Analysis.	Institutions of different sizes may have varying affordability thresholds for medical AI applications, and BIA helps to assess affordability on the institutional level. An assessment on the local level is essential if it intends to inform decision-making locally. The assessment furthermore strongly relies on the local context in which the AI technology might be embedded. This translation is vital, as the general assessment of an AI technology does not necessarily translate towards the local context, calling for a different perspective

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Continuous learning & drift	Technical	Faster updates can cause performance can change post-launch	**Version ID & change log**; drift metrics; plan for monitoring continuous learning.	Version-bound decisions; scheduled re-tests; re-review after substantive change.

				Ongoing monitoring; re-validation; rollback/downtime.
Clinical validation & generalisability	Technical	Need broader, multi-site validation and transportability to the intended settings and populations.	Prospective external validation across sites/equipment; subgroup analyses.	Time-limited recommendations pending real-world results; site acceptance testing. In HTA models cost the impact on utilisation, time-to-diagnosis; re-test costs.
Interoperability & data quality	Technical	Wider integration and diverse data standards vs narrow tools.	Interoperability conformance; data lineage/annotation quality; local integration plan.	Gate by site readiness; phased rollout; evidence 'expiry' if context changes. In HTA models cost integration/IT costs; annotation/ queue/throughput effects.
Transparency & explainability	Technical	Opaquer models and system impact require clearer limits, failure modes, logic summaries.	Usability/communication artefacts; explainability where safety/fairness at stake.	Embed on-screen guidance; verify comms in UI/training; refresh with updates. In HTA models cost content/UX updates; training time; incident comms.
Safety evidence & incident handling	Technical	Scarcity of RCTs; evolving risk profile as models update.	Safety plan; Adverse event taxonomy; continuous post-market safety KPIs.	Incident dashboards; pause/rollback triggers; trigger-based re-review. In HTA models cost incident probability/impact; remediation/downtime.

Organisational readiness & workflow	Procedural	Wider workflow change and workforce re-education	Straining plan; workflow impact/throughput measures.	Phased implementation; periodic implementation reviews. In HTA models cost training/change-management; productivity (learning curve).
Ethics, equity, accountability	Procedural	Broader data and system effects increases risk of unfairness and unclear accountability.	Fairness metrics; consent/privacy approach; accountability assignment.	Equity KPIs with thresholds; clarify roles; re-review if gaps emerge. In HTA models capture equity relevant outcomes and sub groups, and mitigation programme costs.
Economics & affordability (CEA)	Tech and Procedural	Governance and refresh cadence enlarge recurring costs; local affordability varies.	Full cost map (infrastructure, **cyber**, re-training, maintenance); local Budget Impact Analysis; update schedule scenarios.	‘Living’ economic model; scenario tests for cadence/scale; local contextisation. In HTA models capture costs related to governance/audit; compute; scale economies; site-level affordability.

Study: Applications of artificial intelligence and the challenges in health technology assessment: a scoping review and framework with a focus on economic dimensions (Health Economics Review, 2025)

Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
Applications of AI & challenges in HTA: scoping review with economic focus; 2025 (Health Economics Review)	Scoping review and framework (economics-oriented)	No 'frontier AI' term. Paper spans Machine learning, deep learning, Generative AI and AI used inside HTA.	AIs using electronic health records (EHRs), medical claims, and real-world**.** **LLM assistants** in SLR/modelling.	**Dynamic tech and comparators, heavier data/governance footprint, less** transparency, data quality/access **issues and skill/gap issues**	**AI-ready data** infrastructure, governance frameworks, collaboration/training) **Applied CEA**: full implementation costs), dynamic CEAs with version/comparator updates.	One study explored the role of AI in diabetes diagnosis and prognosis, emphasizing the use of real-time data collection and predictive modelling. The method utilizes non-invasive screening techniques, offering scalable and cost-effective solutions for underserved populations, in line with HTA objectives of enhancing equity and accessibility.

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Data infrastructure and linkage	Procedural	Frontier systems need **multimodal, high-volume linkage** (Electronic Health Records/claims/real world data) with stronger governance than narrow ML.	Data standards map; linkage plan; IG approvals; compute/security capacity statement.	Pre-appraisal data-readiness check; staged data access; periodic IG review. Data curation, secure compute/storage, governance overhead. Phase-in rollouts with training gates; refresher cycles post-updates. Training/change-management and planning: productivity (AI learning curve) trajectories.
Dynamic economic evaluation	Tech and Procedural	**Rapid version cycles** and evolving **comparators** make ICERs age quickly vs fixed tech.	Version IDs/timelines; comparator evolution plan; evidence freshness/expiry date.	Time-limited recommendations; scheduled CEA refresh (e.g., 6–12 months) or on version change. Re-build/validation effort; update scedule scenarios; decision delay costs.

Use of AI/LLMs inside HTA (SLR/modelling)	Technical	LLM-assisted reviewing/modelling adds **Quality cCntrol risk** (hallucination, generalisability).	Human-in-the-loop SOPs; prompts/configs; replication logs; QC checklist.	Accept AI-assisted outputs only with documented Quality Control; sample audits. Quality Assurance time vs time-savings; rework
Bias, transparency, accountability	Technical	Broader data and general-purpose behaviours increase **bias** risk and accountability ambiguity vs narrow tools.	Bias/fairness analyses; transparency on data lineage/limitations; accountability assignment.	Equity KPIs with thresholds; governance sign-offs; trigger re-review on gaps. Include as HTA costs mitigation programme & monitoring costs.
Access, data quality & governance	Procedural	Frontier use cases need **sustained data access** and high-quality annotations (for human users to understand outputs) at scale.	Dataset documentation; audit schedule; access agreements; quality metrics.	Make data/Quality Control audits a condition of acceptance; periodic audits. Annotation/audit costs; data licensing.

Study: What Makes Artificial Intelligence Exceptional in Health Technology Assessment? (Frontiers in Artificial Intelligence, 2021).
Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
What Makes AI Exceptional in HTA?; 2021 (Frontiers in AI)	Systematic review of challenges; 45 papers (2016–2021)	No 'frontier' term. Paper argues AI Health Technology **exceptionalism** across 5 dimensions: distinctive features (fixed vs adaptive features), systemic impacts, heightened expectations,	Cross-cutting imaging/triage tools; clinician decision aids	**Faster updates, wider system effects, opacity & update problem** (evidence goes stale quicker; tougher withdrawal; bigger ethics/ equity/governance load than fixed, task-specific AI).	Treat adaptive/opaque AI as **system transformers**: bind evidence and recommendations to version/state; plan monitoring, ethics/equity guardrails, and priced update governance. That is, lifecycle HTA; cooperation	In fact, some AIHTs have already been approved by the FDA, such as AI-powered devices to diagnose eye diseases (Samuel and Gemma Derrick 2020). Risks and harms of AI in healthcare are described at all levels, from the clinical encounter (e.g., adverse effects of an AIHT that can spread to entire patient populations, inexplicability of an AI-based medical decision, issues with assigning responsibility for adverse

		ethical issues, and new evaluative constraints. The paper mentions **adaptive, general-purpose, opaque models (which gits the definition of Frontier AI)** needing lifecycle oversight.			with regulators; intensified post-market scrutiny). In a**pplied CEA** (account for update/rollback and system-level consequences).	events, and patients' loss of trust in their provider) to society as a whole (e.g., furthering inequalities due to algorithm training on biased data) (Sparrow and Hatherley 2019). Interestingly, one indication that current HTA processes are not yet well adapted is the fact that a significant number of AIHTs are benefiting from regulatory fasttrack and do not undergo HTA review, a situation that is particularly noticeable in the United States.

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	**Challenge type (Technical / Procedural)**	**What's different for frontier AI**	**Evidence required for HTA**	**Process action (monitoring/review schedule)**
Fixed (locked) vs unlocked/adaptive state	Technical	Adaptive ('unlocked') systems **change post-approval**; locked can	Declare locked/unlocked; version ID & change	Version-bound recommendations; re-test on each substantive

		become **outdated. B**oth risk misalignment with current care.	log; update plan; before/after performance on same test set.	change; stop/go gates. Re-validation & rollback costs; downtime; update fees (‘update problem’).
System-wide effects & tropism	Technical	Frontier tools can **reframe practice** and propagate biases at scale (e.g., AI can have tropism effects on the healthcare system that may shape certain practices and expectations that are not necessarily accepted or cost-effective. An example of this would be an AI trained on medico-administrative data in a context where physicians have often modified their billing to enter the highest paying codes for clinical procedures, causing the algorithm to infer that these codes represent the usual, standard, or common practice to be recommended.)	Pathway-level impact analysis; bias source documentation; utilisation/threshold shiftss.	Stage rollouts; monitor pathway changes; trigger re-review on unintended shifts. Downstream utilisation & costs; over/under-diagnosis scenarios.

Real-world performance & transportability	Technical	New constraints appear at the clinical level because of the greater variation in AI performance between the test environment and the real-word context than those of drugs and medical devices.	Prospective external validation; multi-site evidence; subgroup/setting sensitivity.	Time-limited recommendations pending Real World Evidence; site acceptance testing. Re-testing; heterogeneity handling; capacity/queue effects.
Safety & incident profile	Technical	New failure modes; **harder withdrawal** and visibility than tangible devices/drugs.	Safety plan; Adverse events taxonomy; incident logs; human-in-the-loop checkpoints.	Incident dashboard; pause/rollback rules; mandatory reporting. Incident probability/impact; remediation & service interruption costs.
Ethics, equity, accountability	Procedural	Scale and opacity amplify **inequity** and responsibility; **automation bias** heightens misuse risk.	Equity-split performance; fairness plan; role/accountability matrix; risk communication.	Equity KPIs with thresholds; communication standards; re-review on gaps. Include in HTA mitigation program & communication costs. DCEA.
Regulatory/HTA infrastructure maturity	Procedural	**Under-developed** AI-ready processes mean more reliance on **lifecycle HTA** vs once-and-done reviews.	Evidence plan for lifecycle; cooperation plan with regulator; transparency of model state.	Scheduled refresh (e.g., 6–12 months) and on change; joint regulator-HTA checkpoints. Ongoing governance/audit; assessor time; Value Of

				Information for re-evaluation cadence.

Study: Value assessment of artificial intelligence in medical imaging: a scoping review (BMC Medical Imaging, 2022).
Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
Value assessment of AI in medical imaging: scoping review; 2022	Scoping review of 86 papers; proposes **11**	No ‘frontier AI’ term. Highlights **explainability, data**	**Large Language Model/Machine**	For frontier-style imaging models: **faster evidence expiry**, bigger **legal/ethical**	**T**ailored HTA that includes algorithm/validatioand	Not stated. Most of the published AI studies within medical imaging

	domains and notes most papers voice needs rather than report outcomes.	**quality/annotation, external validation, integration** which are all amplified in **Frontier AI**	**Learning diagnostic assistants** (radiology, radiomics, oncology, ophthalmology), decision-support in imaging pathways.	footprint, need for **multi-site Real World Evidence** and versioned validation. Few true evaluations; gaps in clinical utility/safety evidence	legal/ethical domains. **Applied CEA** include workflow/time, biopsies avoided.	are retrospective with a technical focus, including reporting of clinical performance metrics, validation, or robustness of the model.

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Algorithm development, performanceand validation	Technical	Needs **multi-site external validation** and version-linked results beyond fixed AI.	**External site results**; reference standards; annotation quality; version ID.	Bind decisions to version; re-test on substantive updates. Re-validation cycles; dataset curation/annotation.
Technology aspects (data quality, IT integration)	Technical	Scale & multimodality increase **black-box** risk and integration load vs narrow AI.	Explainability limits; data lineage; IT integration plan.	Site acceptance tests; staged roll-out. Costs to include in HTA model: Interface work; IT

				integration; downtime/throughput effects.
Clinical effectiveness	Technical	Evidence often **absent**; frontier AI need **utility beyond accuracy and accuracy** across sites/subgroups.	Prospective/real-world utility outcomes (diagnostic outcome, unnecessary biopsies avoided).	Time-limited recommendations pending Real World Evidence. Downstream utilisation; harm/benefit scenarios.
Economics	Procedural	Frontier adds **recurring** costs (monitoring, updates) and system-level effects vs fixed AI.	Workload/time savings, avoided procedures; full cost map (IT, training, maintenance).	Require CEA and Budget Impact Analysis with refresh on version change. Staff time, biopsy reduction, re-training & maintenance.
Ethical / Legal	Procedural	Heightened **privacy, consent, liability** and ownership questions at scale.	Data ownership & consent approach; privacy/security controls; liability allocation.	Governance sign-offs; legal review gates. Costs to include in HTA model: Compliance overhead; insurance/indemnity exposure.
Organisational & Patient/Social	Procedural	Larger **workflow redesign** and acceptability risks for general-purpose models.	Workflow/time studies; clinician & patient acceptability metrics.	Phased implementation with training gates. Costs to include in HTA model: training/change management; capacity/queues.

Study: The Intersection of Digital Health and Artificial Intelligence: Clearing the Cloud of Uncertainty : *Digital Health*, 2025.

Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
The intersection of digital health and artificial intelligence: clearing the cloud of uncertainty; 2025	Commentary / conceptual policy paper	No explicit 'frontier AI,' but introduces **LLMs and generative AI** as emerging subtypes requiring tailored regulation and HTA; maps well to **'frontier' models.**	**LLM diagnostic/administrative assistants, AI-enabled Digital Health Technology** (mHealth, wearables, electronic medical records-integrated apps).	Frontier AI terminological confusion (differences between digital health vs AI), governance/ethical oversight gaps, adds **bias, hallucination, and data drift** concerns; demands **continuous validation** and explicit management of ethical and performance risks beyond traditional AI.	Category-specific frameworks, Real World Data quality standards, bias/validation protocols. Clear conceptual distinction between digital health and AI is foundational to credible HTA and risk-based frameworks for AI-enabled technologies.	AI can be embedded in health technologies, such as medical devices, wearables, mobile applications, or web platforms, to enhance their functionality and performance or to leverage collected data to advance precision and personalized medicine. AI tools have the potential to reduce human errors because of their ability to decrease human workload and empower individuals to think and perform at a higher level of quality.

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Definition & categorisation	Procedural	Frontier AI blurs DH/AI boundaries; unclear categorisation hampers HTA scope.	Category definition (Digital health techology vs Digital health service vs AI tool) and function mapping.	Require early scoping checklists in appraisals. Administrative complexity/time for reclassification.
Algorithmic transparency & bias	Technical	Generative/LLMs raise new bias, misinformation, and opacity issues.	Model card (data sources, bias mitigation, drift management).	Continuous audit post-market. Include in models cost of monitoring bias & retraining.
Data & RWD integration	Technical	Frontier AI draws on vast, heterogeneous Real world data; data lineage essential.	Provenance, representativeness, quality criteria.	Link data-quality assurance to reimbursement eligibility. Real World Data curation/infrastructure costs.
Ethical, legal, and privacy	Procedural	Frontier AI amplifies consent, IP, and accountability concerns.	Ethics plan, privacy safeguard, ownership statement.	Governance review and sign-off. Include in models compliance and liability costs.
Clinical & economic evaluation	Technical	Dynamic learning systems complicate static CEA models.	Update frequency, version control, outcome linkage.	Lifecycle HTA with periodic refresh. Incorporate in models re-validation and lifecycle costs.

Study: Recommendations to Overcome Barriers to the Use of Artificial-Intelligence–Driven Evidence in Health Technology Assessment (Frontiers in Public Health, 2023).

Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
Recommendations to Overcome Barriers to the Use of AI-Driven Evidence in HTA (2023)	Original research paper based on survey and expert workshop.	No explicit 'frontier AI' definition; focus on AI/ML for **evidence generation** within HTA. Maps to 'frontier AI used for evidence production and policy analytics.'	Administrative and analytical GenAI / HTA analytics assistant / AI-driven evidence synthesis tools.	Frontier models increase demand for **governance, data standardization, validation** capacity and shared AI infrastructure; they stress data infrastructure and skills gaps. Frontier AI can be used to generate real-world evidence	education, collaboration, data standards, governance. **CEA:** mentions Value of Information analysis and economic assessment of data validity investments.	Not reported. Sophisticated computational models and algorithms, combined with powerful computers and the availability of vast amounts of data, have recently accelerated the application of artificial intelligence in various fields. This is primarily driven by the development of machine learning, which has contributed to advances in data science and statistical prediction.

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Human factors / capacity	Procedural	Requires cross-disciplinary AI literacy and shared expertise across HTA 'doers' and 'users.'	Training plans, course completion metrics.	Institutional capacity plans; education tracking. Include in HTA models the cost of training & capacity development.
Governance & policy frameworks	Procedural	Frontier AI needs political commitment and regulatory clarity for sensitive data sharing.	Evidence of national digitization policies and database regulations.	Include policy alignment check in HTA protocols. Include in HTA models the cost of delayed adoption vs benefits of regulatory enablement.
Data quality & standardization	Technical	Frontier AI intensifies cross-border data integration issues	Data mapping documentation; CDM adoption proof.	Mandate standard data models and federated approaches. Include in HTA models the value of information analysis for data improvement.
Bias & validity	Technical	Bias amplified by complex models and unequal representation in training data.	Bias audit and sensitivity analysis plans.	Iterative validation and auditing requirements. Include in HTA models the downstream impact of biased outputs on CEA parameters.

Infrastructure & resources	Procedural	Requires sustainable AI computational capacity (shared centres of excellence).	Infrastructure plans and funding models.	Mandate reuse of existing platforms; sustainability reporting. Capital vs operational cost allocation for AI infrastructure.
Transparency & reporting	Technical	AI tools must be **reproducible and explainable** for HTA documentation.	Lay summaries of AI algorithms; training data description.	Require transparent reporting in HTA appraisals. Include in HTA models the cost of non-replicable analysis (error and revision risk).

Study: Capabilities and risks from frontier AI: A discussion paper on the need for further research into AI risk (UK, Oct 2023)
Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
Frontier AI: capabilities and risks – discussion paper (DSIT; 2025)	UK government discussion paper (research	Defines 'frontier AI' as 'highly capable general-purpose AI	General-purpose LLM/LMM foundation models powering:	Open-ended behaviour, emerging autonomy, fast post-training upgrades (tools,	External assurance/monitoring; safer deployment choices.	Not stated. Defining AI is challenging as it remains a quickly evolving technology. For the purposes of the Summit we define 'frontier

	& analysis).	models that can perform a wide variety of tasks and match or exceed the capabilities present in today's most advanced models'; today this primarily includes LLMs, noting future systems may use other tech.	clinical copilot, multimodal diagnostic assistant, admin GenAI, and autonomous agents/scaffolds (e.g., AutoGPT-style).	prompts, scaffolds), lack of robust safety standards, black-box interpretability, non-robustness risk; therefore higher monitoring burden vs. non-frontier AI.	Include in the model: dynamic performance change; system-level effects (efficiency/productivity). HTA should assume higher evidence and surveillance needs than for fixed, narrow AI.	AI' as highly capable general-purpose AI models that can perform a wide variety of tasks and match or exceed the capabilities present in today's most advanced models (seeFigure 1).Today, this primarily includes large language models(LLMs)such as those underlying ChatGPT Claude,and Bard. However, it is important to note that, both today and in the future, frontier AI systems may not be underpinned by LLMs, and could be underpinned by another technology.

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Clinical effectiveness	Technical	Open-ended, general-purpose behavior; performance can alter with	Task-specific prospective evaluations with pre-registered	Early conditional adoption with rapid evidence updates

		tools/prompts/scaffolds (Researchers have built software programs called ‘scaffolds’ that allow frontier AI models to power autonomous AI agents..	endpoints; sensitivity to prompt/tool/scaffold variants.	(e.g., 6–12-month review). Include in model the AI model learning curve; benefit decay/gain under new tools/prompts.
Safety/harms	Technical	In general, frontier AI systems are not robust, i.e. they frequently fail institutions sufficiently unlike their training data.; hallucinations.	Adversarial tests; failure taxonomy and rates in clinical-like tasks.	Pre-launch external assurance; post-market safety signal surveillance. Include in model cost of adverse events; mitigation program costs.
Generalisability/external validity	Technical	Performance sensitive to distribution shift and context length/long-horizon tasks.	Multi-site, real-world validation; stress tests on re-worded/altered inputs.	Triggered re-evaluation when indications/data domains change. Include in model scenario analysis for alternative risk; value of information for more sites.
Transparency/interpretability	Technical	‘Black-box’ parameters; immature mechanistic interpretability.	Documentation of model lineage, data classes, post-training steps; interpretable error analyses.	Require model cards/assurance packages at appraisal and on version change. Include in model cost of explainability

				tooling vs. decision quality gains.
Bias & equity	Technical	Risk of inherited/amplified biases from web-scale data and open-ended outputs.	Stratified performance/bias audits across protected groups and settings.	Equity impact review at fixed intervals; corrective fine-tuning plans. Distributional cost-effectiveness; inequity weights.
Cybersecurity & misuse	Technical/Procedural	Elevated misuse surface (prompt-injection, agentic autonomy); offence–defence arms race.	Threat modelling; penetration tests including prompt-injection; secure deployment evidence.	Continuous monitoring; incident reporting requirements. Include in model expected loss from breaches/downtime; security control costs.
Data governance & privacy	Procedural	Tool-use/agents may access broader data; harder provenance tracking.	Provenance logs; access controls audit; data-minimisation proof.	Audit logs and periodic compliance attestations. Include in model compliance/guardrail costs vs. penalties avoided.
Real-world performance monitoring	Procedural	Capabilities can improve (or degrade) post-deployment with fine-tuning/scaffolds.	KPI pack: accuracy, hallucination rate, escalation rate; version/change logs.	Rolling Real World Evidence collection; 6-monthly surveillance with trigger rules. Include in model cost of surveillance;

				benefit from faster learning cycles.
System/implementation impact	Procedural	Large productivity effects beyond clinical end-points (workflow, admin).	Time-motion studies; queue and throughput metrics; staff acceptance.	Phased rollout with operational dashboards. Include in model downstream staffing/time savings; induced demand.
Versioning & change control	Procedural	Frequent model updates can change AI behavior.	Semantic versioning; change logs; re-test on release.	Require re-appraisal on major change; light-touch check on minors. Include in model re-validation costs; downtime and switch-over overheads.

Study: Draft Report of the Joint California Policy Working Group on AI Frontier Models (March 18, 2025).

Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
Draft Report of the Joint California Policy Working Group on AI Frontier Models (2025)	State policy **draft** report; evidence-based governance framework	Defines **frontier models** as the most capable subset of foundation models; cites March 2025 examples (o3, Gemini 2.0, Claude 3.7 Sonnet, Llama 3.3, etc.).	General-purpose **LLM/LMM foundation models** underpinning multiple archetypes (clinical copilot, GenAI, agentic tooling). The report is sector-agnostic but notes economy-	Highlights **rapidly evolving capabilities**, **evidence gaps**, **transparency deficits**, need for **independent/third-party evaluation**, and **post-deployment adverse-event reporting**, all heavier burdens than for narrow/non-frontier AI. Other challenges ate defining reportable events, under-reporting,	Guiding principles for transparency and adverse event reporting.	Non stated. Improvements in capabilities across frontier AI models and companies tied to biology are especially striking. For example, OpenAI's o3 model outperforms 94% of expert virologists. OpenAI's April 2025 o3 and o4-mini System Card states, 'As we wrote in our deep research system card, several of our biology evaluations indicate our models are on the cusp of being able to meaningfully help novices create known biological threats, which would cross our high risk threshold.

			wide implication.	monitoring costs, updating thresholds.		We expect current trends of rapidly increasing capability to continue, and for models to cross this threshold in the near future.’

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What’s different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Clinical effectiveness	Technical	Emphasises **independent/third-party assessments** due to immature/fast-moving evaluation.	For health use, minimally: task-specific third-party evaluations and disclosure of evaluation methods (mapped from transparency principles).	Require re-testing when capabilities, use context, or thresholds (thresholds for policy interventions, such as for disclosure requirements or diagnostic accuracy thresholds) change. Scenario costs for repeated evaluations.
Safety/harms	Technical	Advocates **adverse-event reporting (AER)** to learn realised harms/unanticipated risks.	AER schema: clear event definitions, data fields, reporter classes, recipient agency.	AER (mandatory developer and voluntary user), with periodic criteria updates. Cost of AER reporting and impact of detected incidents.

External validity	Technical	Notes **evidence gaps** for rapidly evolving models; stresses transparency & third-party checks.	Independent replication on representative settings	Trigger re-review when deployment alters or new risks observed via AER. Budget for repeat testing across sites.
Transparency	Procedural	Calls for **improving transparency** (capabilities, mitigations, documentation) balanced with security.	Minimum: model/system documentation sufficient for external assessment (as per guiding principles).	Periodic transparency updates aligned to threshold changes and AER insights. Model cost of transparency/reporting programs.
Bias & equity	Technical	Not developed in health terms.	Require stratified reporting where applicable.	Fold into AER review cycles. Potential equity impact analyses in CEA
Cybersecurity & misuse	Technical/Procedural	Uses **incident/AER** framing and information-sharing analogies.	Incident definitions and sharing pipelines where relevant.	Continuous monitoring; periodic criteria revision. Model expected-loss from incidents; controls spend.
Real-world performance monitoring	Procedural	**AER and thresholds** act as post-deployment monitoring infrastructure.	Minimal KPI set drawn from AER fields; clear trigger thresholds.	Regular analysis of AER data; scheduled threshold updates. Include in models ongoing monitoring costs.
Versioning & change control	Procedural	Recommends **updatable thresholds** as tech/society evolve;	Documented change logs and re-	Periodic threshold review; trigger rules

		implies re-assessment on major model shifts.	assessment against current thresholds.	when metrics breached. Model re-validation costs/downtime.

Study: Digital Transformation Needs Trustworthy Artificial Intelligence (Mayo Clinic Proceedings: Digital Health, 2023)
Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
*Digital Transformation Needs Trustworthy Artificial Intelligence (*2023)	Editorial / viewpoint (medical journal).	No ‘frontier AI’ definition; references ChatGPT-4 as an illustrative modern ML system.	General-purpose **LLM** as example; not tied to a specific healthcare archetype (trust, transparency/interp	None stated; arguments are model-agnostic principles (trust, XAI, ethics).	Human in-the-loop; requirements for safety, traceability, transparency, explicability, validity, verifiability. **Applied CEA:** not covered.	Not stated. Perhaps the most important topic of AI in medicine, but also in many other areas of application, is trust. The new work by Farah et al shows very impressively and convincingly that trust in AI-based medical devices depends on transparency

			retability, explainability/ ethical/legal analysis; context-/risk-dependent standards.)		High-level call for trustworthy, explainable medical AI	(interpretability and explainability of the results) and ethics (in the sense of trustworthiness and regulation)

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Clinical effectiveness	Technical	Emphasis on **explainability alongside accuracy**; standards context/risk-dependent.	Report task metrics (e.g., rationale/feature-attribution) sufficient for clinician review.	Revisit evidence when indication/risk changes. Potential trade-off between explainability and performance.
Safety/harms	Technical	Trust depends on demonstrable **safety and** verifiability	Document failure modes; robustness checks; human-in-the-loop escalation paths.	Post-deployment safety checks aligned to risk class. Model cost of safety mitigations vs avoided adverse events.

Transparency/traceability	Procedural	Calls for **traceability** (data/method lineage) to underpin trust.	Basic model card: data classes, training/validation approach, known limits.	Update documentation on version change. Include in model documentation/programme costs.
Interpretability/explainability	Technical	**Explainability** positioned as cornerstone of trust.	Provide clinician-interpretable outputs or explanations appropriate to use case.	Periodic usability audits of explanations with clinicians. Model impact of explanations on decision quality/workflow time.
Ethics & governance	Procedural	Ethical assurance (validity, verifiability) highlighted. Trustworthiness implies **fairness** considerations.	Statement of ethical controls; verification steps for claimed use.	Include ethics checkpoint in governance gates. Model governance overheads.

Study: Acceptance of Artificial Intelligence in Evidence and Dossier Developments by HTA bodies: Challenges and Opportunities (Putnam, 20 May 2025).

Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
Acceptance of AI in Evidence & Dossier Development by	Consultancy blog/insight post	No 'frontier AI' term; focuses on **LLMs/ML** for	enabled efficiency across evidence	Stresses **transparency/replicability** vs. black-box AI;	Publish HTA AI guidance; third-party/independent checks;	Non stated. Artificial Intelligence (AI) holds transformative potential in healthcare,

HTA bodies (Putnam; 2025)	(thought leadership).	evidence generation and HTA workflows.	workflows. assisted **SLR/screening**, **data extraction and synthesis**, **Real World Evidence support**, and **assistance in economic model development**	risks: **hallucination**, **bias**, **provenance**, **privacy** implying higher disclosure/audit needs than traditional methods.	pilots; training; human oversight. **Explicit standards, pilots, and oversight**.	particularly in the field of Health Technology Assessment. By leveraging machine learning, natural language processing and especially large language models, AI algorithms can significantly expedite evidence generation by analyzing vast and unstructured datasets with greater efficiency and accuracy.

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Clinical effectiveness / evidence generation	Technical	Can **accelerate literature reviews, extraction, synthesis** but risk i**naccuracy/hallucination**.	Declare AI use; locked protocols; independent test/Quality Assurance of AI-assisted outputs; sensitivity analyses vs. manual baselines.	Quality Assurance for AI-assisted steps; re-check when models/prompts change. Include in models cost/time savings from AI vs. rework from errors;

				Value Of Information of extra validation.
Methods transparency & reproducibility	Procedural	HTA’s bar for **replicability/provenance** clashes with black-box models.	Method card: data sources, prompts/workflows, human oversight points, audit logs.	Require declaration in appraisals; audit trail review each update. Documentation/audit program costs vs. acceptance probability.
Bias & equity & reliability of AI-supported evidence	Technical	**Bias amplification** possible from training data and retrieval.	Stratified error/performance checks; provenance of sources; bias mitigation steps. Report discrepancy rates and corrective actions.	Periodic bias audits when topic/population alters. Distributional CE (equity impacts) from biased evidence. Pre-appraisal pilot then scale; ongoing spot checks.
Data governance & privacy	Procedural	Use of **patient-level data** raises confidentiality risk.	Data minimisation; de-identification proof; access controls for AI tools.	Compliance attestations per appraisal; incident reporting channel. Include in models compliance/guardrail costs; risk-adjusted penalty avoidance.
Human oversight	Procedural	Article argues for **hybrid** (AI and human expert oversight) workflows.	Sign-off points by qualified reviewers; discrepancy adjudication logs.	Embed reviewer checkpoints per major AI use. Include in models reviewer time vs. error reduction; productivity gains.

Pilots & capability building	Procedural	**Pilots** and **training** of HTA assessors.	Documented pilot case studies with predefined success criteria.	Annual training plan; post-pilot lessons-learned cycle. Include in models the cost of pilot/programme costs vs. future efficiency/acceptance benefits.

Study: Foundation models for generalist medical artificial intelligence (GMAI) : *Nature* Perspective, April 2023.
Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
Foundation models for generalist medical AI; 2023 (Nature)	Perspective proposing **Generalist Medical AI** built from **foundation models** (self-supervised,	Doesn't use 'frontier AI' term, but explicitly defines **foundation models/GMAI** with emergent capabilities (in-context	**Large language model clinician copilot**, **multimodal diagnostic assistant**, **bedside**	**Validation & verification are harder** due to open-ended tasks; **social bias, scale/compute, environmental cost** and **version drift** exceed typical ML;	Explainability, uncertainty expression, audits; lifecycle oversight. **Applied CEA** include costs of scale/compute, monitoring, incident management.	Not stated. Although there have been early efforts to develop medical foundation models, this shift has not yet widely permeated medical AI, owing to the difficulty of accessing large, diverse medical datasets, the complexity of the medical domain and the recency of

	multimodal, in-context learning) with dynamic task specification and formal medical knowledge).	learning, multimodal I/O, medical reasoning). This maps directly to **frontier** per characteristics (general-purpose, high-compute, rapidly updating).	**early-warning and recommendations**, **patient chatbots**, **interactive note-taking**.	require **new governance and lifecycle HTA**.	Tech needing **version-bound approvals, strong transparency, Real world evidence, and priced governance**.	this development. Instead, medical AI models are largely still developed with a task-specific approach to model development. For instance, a chest X-ray interpretation model may be trained on a dataset in which every image has been explicitly labelled as positive or negative for pneumonia, probably requiring substantial annotation effort. This model would only detect pneumonia and would not be able to carry out the complete diagnostic exercise of writing a comprehensive radiology report. This narrow, task specific approach produces inflexible models, limited to carrying out tasks predefined by the training dataset and its labels. In current practice, such models typically cannot adapt to other tasks (or even to different data distributions for the same task) without being

						retrained on another dataset. Of the more than 500 AI models for clinical medicine that have received approval by the Food and Drug Administration, most have been approved for only 1 or 2 narrow tasks.

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Intended use and scope control	Procedural	Open-ended, dynamically specified tasks. Harder to **anticipate failure modes** across unseen tasks; verification may need **multidisciplinary panels**.	Declared **approved task set**; guardrails for unsupported tasks; interface warnings for off-label prompts.	Version-linked approvals; User Interface-level guardrails; stop/go criteria. Time-limited recommendation pending Real World Evidence; periodic stress-tests; challenge audits.

				Include in models Governance & User Interface content; incident handling for off-label use. Panel time; repeat testing; Value Of Information for re-assessment cadence.
Bias and equity	Technical	**Bias can grow with model scale** and propagate widely across downstream applications.	Subgroup performance; bias audits; mitigation plan & thresholds.	Equity KPIs; re-review on gaps; publish subgroup dashboards. Distributional CEA; mitigation programme costs.
Privacy and security	Technical	Data extraction and **prompt attacks** more salient vs fixed ML.	de-ID. Approved deployment patterns; periodic privacy red-teaming.	Secure hosting/incident reporting; additional compliance. Include in models costs of possible data security breach.
Scale/compute and environment	Procedural	**High compute/** training/inference; potential **environmental footprint**.	Compute/energy profile; Efficiency targets	Energy/hosting costs; refresh cycles; carbon scenarios.
Reasoning and explainability	Technical	Free-text reasoning and multimodal justifications need **fact-checking** aids.	sample fact-checks per cycle.	Model Quality control time vs errors avoided

Pathway impact	Tech and Procedural	Broad use cases can lead to **system-level** effects (utilisation alters, task altering).	Pathway maps; utilisation thresholds; human-in-the-loop points.	Stage roll-out; monitor demand spillovers; triggers for review. Model queue/capacity; downstream procedures; staff mix changes.

Study: INAHTA Position Statement: Disruptive Technologies : International Network of Agencies for Health Technology Assessment (Position Statement), March 2022.

Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
INAHTA Position Statement: Disruptive Technologies; 2022	Position statement / policy	No ‘frontier AI’ term. Uses **Christensen** definition: disruptive tech is *cheaper, simpler, more*	Large language model **admin scribes/agents** (task shift to lower-cost	Frontier AI is **more likely to be pathway- and business-model-disruptive** (task altering, site altering, standardisation)	Include in HTA organisational and economic; lifecycle surveillance; adoption mechanisms **Applied CEA** include	Identification of disruptive technologies is often complicated by the fact that disruptiveness can often only be judged in retrospect, after markets have changed. For the diagnosis and therapy of acute conditions

		mobile, potentially higher quality, can be delivered by **differently skilled providers**, and **displaces** incumbents via a new **business model**; market uptake is the key signal. Similar to Frontier AI in that it can be **general-purpose, rapidly scaling AI that can reconfigure workflows/providers**.	staff), Large language model **triage/diagnostic assistants**, **remote monitoring** models changing care location	Identification of disruptive tech is ex ante, data paucity, need for Real world Evidence	adoption instruments & business-model alters. If AI looks **disruptive**, plan a **comprehensive, lifecycle HTA** and **price adoption mechanisms**, with ongoing Real World Evidnce to confirm net value.	(e.g., rapid antigen testing for streptococcal pharyngitis), disruptiveness could be associated with the involvement of differently skilled and equipped providers. For chronic disease, disruptiveness could be associated with a different management model (e.g., care management through nurses and remote monitoring in patients with heart failure).

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Disruptiveness screening	Procedural	Frontier AI more often meets **predictors** (cheaper/faster, point-of-care, task alter to lower-skilled providers, pathway change).	Short 'predictors' checklist; anticipated pathway and provider changes; access/throughput effects.	Flag as 'potentially disruptive' then escalate to **comprehensive** HTA. Scenario analysis for **task altering**, site altering, and throughput; scale-up curves.
Organisational impact	Procedural	Frontier tools can **reallocate roles** (e.g., from specialists to nurses/assistants) and move care from hospital to community.	Pathway map; staffing/skills model; facility readiness.	Include organisational domain formally; stage rollouts with acceptance gates. Workforce/training; capacity/queue effects; facility reconfiguration.
Economic consequences	Tech and Procedural	Business-model change (lower unit cost, higher volumes) makes **budget impact** and **price/volume** dynamics central.	Unit cost vs incumbent; expected volume shift; acquisition vs operating mix.	Require **Budget Impact Analysis + CEA** with sensitivity to uptake, displacement, and price evolution. Price/volume scenarios; displacement of legacy tech; learning-curve effects.
Identification timing & uncertainty	Procedural	True disruptiveness often visible **only in retrospect**; frontier AI diffuses quickly leads to higher ex-ante uncertainty.	Early **World Evidence** plan; leading indicators (market share, site count)	**Coverage with Evidence Development** or time-limited approvals with **World Evidence** checkpoints. Include in models the cost of

				registries, audits; risk of wrong-way adoption.
Real World Evidence and lifecycle surveillance	Technical	Faster adoption and pathway change need **post-market Real World Evidence** to confirm net value and equity.	**Real World Evidence** outcomes include utilisation, access, equity; AE/incident logs; market uptake metrics.	Monitor across lifecycle; re-review as pathways evolve; consider delisting inferior tech. Include in HTA models ongoing monitoring costs; de-implementation costs of displaced tech.
Adoption mechanisms	Procedural	To unlock benefits, frontier AI may need **policy levers** (reimbursement pilots, incentives).	Proposed incentives/reimbursement model; conditions and KPIs.	Conditional coverage, targeted incentives, or managed entry with sunset reviews. May need to model cost of incentives; administrative overhead; KPI-linked continuation/withdrawal.

=

Study: 2025 Watch List: Artificial Intelligence in Health Care : Canada's Drug Agency (Horizon Scan), March 2025.

Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
2025 Watch List: Artificial Intelligence in Health Care; 2025	Horizon scan (top 5 AI techs + top 5 cross-	No 'frontier AI' term; defines **Large Language**	**Admin GenAI (AI notetaking/scribes**	Broader **system effects** (e.g. capacity), **faster update cadence**, **unsanctioned**	A policy-ready map of **where AI will hit systems first** (scribes,	Non stated. A recent randomized controlled trial conducted in the US found that the LLM ChatGPT Plus alone

(Canada's Drug Agency)	cutting issues)	**Models** and **AI agents**; scope is AI tech 'that can replace, displace, or augment tasks' with system impacts. Therefore links with **AI agents/general-purpose, rapidly evolving tools**.	); **Large Language Models for training/education aids** and as **diagnostic assistants**; **AI treatment optimisers**; **AI remote monitoring**.	**use risk**, and **environmental costs** leads to stronger transparency, governance and living HTA	training, Diagnostics, Treatments) and **what guardrails** (privacy, liability, data, sovereignty, environment) HTA should price and enforce.	demonstrated higher performance in diagnostic reasoning compared with physicians, even when the LLM was available to them. However, when physicians used the LLM as a diagnostic aid, it did not statistically significantly enhance their clinical reasoning or reduce time spent per case compared with using traditional resources, such as UpToDate or Google

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Admin GenAI (AI notetaking/scribes)	Tech	**Large Language Model hallucination/omission** risk and multilingual variability at scale.	Task accuracy vs human notes; error taxonomy; User Interface review and clinician sign-off.	Time-limited rollout; Quality control sampling; re-test on model/version change. HTA could model the admin time saved vs

				Quality control rework; training and integration costs.
Large Language Model tools for training/education	Procedural	Rapid content drift; dependence on prompt quality; risk of overreliance.	Curriculum mapping; validation of citations; usage/log policies.	Governance for approved use; periodic material refresh. Include in HTA model the cost of Training time offsets; staff productivity effects.
Disease detection/diagnosis	Tech	Earlier Diagnosis but **demand spillovers** and overdiagnosis risks scale faster.	External multi-site validation; thresholds; downstream utilisation modeling.	Stage adoption; monitor workloads & false positives; trigger re-review. Include in HTA models capacity/queue effects; downstream tests/procedures; harm/benefit scenarios.
Treatment optimisation / chatbots	Tech and Procedural	Personalisation and 24/7 agents leads to **liability** & safety governance more complex.	Human-in-the-loop design; safety cases; consent and comms artefacts.	Consent & supervision requirements; incident reporting cadence. Include in HTA models support costs.
Remote monitoring (AI-enabled)	Tech	Continuous data and alerts risk **false alarms** and digital divide inequities.	Subgroup accuracy (e.g., oximetry skin-tone bias); connectivity plan.	Local acceptance tests; equity KPIs; drift monitoring.
Privacy & security	Procedural	**Large Language Models** /agents heighten exposure;	Storage/region; encryption/logging;	Information governance sign-off; periodic security tests;

		physician confidence low.	breach history; opt-out/consent flows.	transparency to users. Include in HTA model costs security; incident response; consent workflows.
Liability & accountability	Procedural	Agentic behaviour/opacity blur **duty of care** lines.	Accountability matrix (developer/provider/user); disclosure policy; human-in-command.	Clinician guidance/labeling; audit trail. Include in models the cost of legal/compliance overhead; insurance; downtime from pauses.
Data quality, bias & availability	Tech	Fragmented data and bias risks create **garbage-in/garbage-out** at scale.	Dataset lineage; representativeness; bias tests by subgroup; interoperability plan.	Data standards; pan-system sharing rules; periodic bias audits. Include in models Data curation/ interoperability costs; mitigation programmes
Data sovereignty & governance	Procedural	Need explicit community control and lawful use across jurisdictions.	Data ownership/consent model; localisation; governance board minutes.	Enforce localisation/sovereignty constraints; stop-go gates. Include in costs local hosting; governance staffing.
Environmental costs	Procedural	Compute-heavy agents add **energy & e-waste externalities**.	Energy profile/hosting; lifecycle plan; environmental impact estimates.	Prefer efficiency targets; greener deployment patterns. Include in HTA models energy/hosting; refresh cycles; carbon pricing sensitivity.

Study: An Overview of Continuous Learning Artificial Intelligence-Enabled Medical Devices : CADTH Horizon Scan (May 2022).

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
Continuous-Learning AI-Enabled Medical Devices; 2022 (CADTH Horizon Scan)	Horizon scan (landscape of **adaptive/continuous-learning** vs **locked** AI; regulatory + HTA implications)	Distinguishes **continuous-learning (adaptive/unlocked)** from **fixed/locked** AI; highlights **non-stationarity, drift, transparency, and lifecycle oversight** needs; no	SaMD/SiMD that update in use (e.g., imaging/diagnostic assistants, clinician decision aids)	**Performance changes after deployment**, larger surveillance burden, harder withdrawal, and faster evidence expiry than task-specific/locked tools. drift, transparency, generalisability, safety.	Treat adaptive AI as **living devices**: tie decisions to version/state and cost the governance and refresh burden alongside benefits (require **version-bound evidence, continuous monitoring, change**	As the majority of locked AI or continuous learning AI that are in use or approved for the market are in the areas of diagnosis and imaging,3 it is likely that these areas will be the first to see approved continuous learning AI. Some have proposed that diagnostic testing would be an area where continuous learning models could be implemented safely.8 For example, the AI could

		‘frontier’ term used. Map **frontier ≈ adaptive/rapidly updating systems**.			**control,** and shorter re-review). Post-market monitoring, stakeholder dialogue, **Applied CEA** include recurring monitoring/rollback costs; dynamic value).	make a prediction, and then the clinicians verify the diagnosis, labelling the data and providing it back to the AI to self-adjust

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What’s different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Versioning and drift	Technical	**Continuous learning** can lead to performance alters post-deployment	Version ID and **change log**; pre/post-update tests on the same set; declared update plan.	**Version-bound** recommendations; scheduled re-tests; trigger re-review on substantive change. Include in models monitoring and re-validation; rollback/downtime scenarios.
Real-world transportability	Technical	Bigger test (due to bigger scope of tasks) means real-world gap;	Multi-site external validation; subgroup results; Real world	Time-limited adoption pending Real World Evidence; site

		heterogeneity across sites/populations.	Evidence plan with cadence.	acceptance testing. Include in models re-test costs; utilisation and pathway effects.
Transparency and explainability	Technical	Adaptive updates can reduce explainability and traceability over time.	Limits/failure modes; logic summaries (where feasible); data lineage.	UI/label updates at high-risk steps; refresh comms on update. Include in HTA models communications/training refresh; incident communications.
Safety and incidents	Technical	New failure modes can **emerge after updates**; withdrawal is harder.	Safety plan; Adverse event taxonomy; incident logs; thresholds.	Live incident dashboard; **pause/rollback** rules; mandatory reporting. Include in HTA models incident probability/impact; remediation downtime.
Governance and change control	Procedural	Needs change protocols; cross-regulator alignment.	Declared change protocol; monitoring metrics; accountability roles.	‘Living’ oversight; regulator-HTA checkpoints; shorter cycles. Include in HTA models ongoing audit/governance; assessor time.
Ethics, equity & accountability	Procedural	Continuous learning can alter responsibility and amplify subgroup bias.	Equity-split performance; consent/privacy approach; role assignment.	Equity KPIs with thresholds; re-review on gaps.

Study: AI in Health: Huge Potential, Huge Risks : OECD policy brief (2024).

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
AI in Health: Huge Potential, Huge Risks (OECD; 2024)	OECD **policy brief** on operationalising trustworthy AI in health (ties to 2019 OECD AI Principles).	No 'frontier AI' term; frames AI via **OECD AI Principles** (human-centred values, transparency, robustness, accountability).	Broad: Large Language Models - enabled clinical & admin tools; decision support; cybersecurity analytics; population health AI.	Elevates **governance/liability**, **workforce disruption/upskilling**, **equity & scale**, **privacy/security**, **transparency/explainability**, and **certification/incident reporting**.	Governance/oversight, metrics and public reporting, liability clarity, workforce training, interoperability, guidance on 'responsible AI', certification/incident reporting. **Applied CEA hooks:** productivity/time savings, cyber risk losses, data infrastructure	None stated Bespoke AI applications without the ability or intention to scale (e.g., due to system incompatibility or lack of technical resources), risk a fragmented set of AI innovations that are built and maintained by wealthy health organisations and only available to wealthy segments of the public.

					costs, equity impacts.	

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Governance, accountability and liability	Procedural	Lack of clear **liability** and oversight for evolving AI; need public reporting of positive/negative **AI incidents**.	Governance plan including roles, liability statement, incident taxonomy and reporting route.	Require incident reporting; periodic governance audits; publish metrics nationally. Include in models expected loss from incidents; cost of governance/assurance.
Workforce impact & capacity	Procedural	AI may automate (paper say up to **~36%** of tasks); skills alter and burnout risks if poorly implemented.	Time–motion baselines; pilot results on task reallocation; training plan.	Phased roll-out with staff feedback; annual skills review. Include in models productivity gains vs. training/change-management costs.
Equity & access / scalability	Procedural	Risk of **bespoke**, unscalable tools that entrench inequity; need population-level availability/impact measures.	Disaggregated uptake/outcome metrics (geography, gender, groups); scalability plan.	Equity checkpoints; cross-site scale reviews. Distributional CEA; scale economies vs. duplication.

Data governance, interoperability	Procedural	Fragmented **policy/data/tech** foundations block scaling and reuse; need alignment with Health Data Governance.	Data lineage, quality controls, interoperability conformance; cross-border data plan.	Data audits; integration milestones; cross-border coordination reviews. Include in models data infrastructure build/maintain costs; benefits from reuse of the 97% ‘unused’ data.
Transparency, explainability & representativeness	Technical	Bias/poor outcomes if data not **representative**; need **explainable** AI to sustain trust.	Model/system card include training data classes, performance by subgroup, explanation artefacts.	Require transparency pack at appraisal and on major updates. Cost of documentation vs. adoption and error reduction.
Privacy & cybersecurity	Technical	Sensitive data and rising **cyber threats** increase privacy/security trade-offs and need for system resilience.	Privacy risk assessment; security testing and breach response plan.	Continuous security monitoring; incident drills; penalties for re-identification. Cost of data breach expected loss; control costs; downtime.
Certification & regulation	Procedural	Market is growing fast with **limited oversight**; need criteria for **‘responsible AI’** and certification paths.	Conformance to defined ‘responsible AI’ criteria; evidence of clinical appropriateness.	Pre/post-market checks; periodic re-certification. Assurance program costs vs. avoided harms and faster uptake.
System outcomes & benefits tracking	Procedural	**Metrics** on adoption, benefits, and harms to	KPI set: adoption rates, workforce time saved,	National dashboards; annual public reports; trigger reviews. Include

		avoid principle–practice gaps.	outcomes, incident counts.	in model net system value from productivity & safety gains.

Study: Alami et al., 2020 : 'Artificial Intelligence and Health Technology Assessment: Anticipating a New Level of Complexity' (JMIR Viewpoint).

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
AI & HTA: anticipating a new level of complexity (Alami et al., 2020)	Peer-reviewed **viewpoint** mapping AI issues onto the **HTA Core Model;** payer perspective.	No 'frontier AI' term; discusses AI broadly (diagnosis, screening, triage, resource allocation).	AI decision support; workflow/triage; admin optimisation; patient-facing follow-up.	Highlights **generalisability gaps, black-box/interpretability, cybersecurity, clinical tropism, workflow burden/automation bias, organisational change, economic investments**, and **legal/ethical liability & consent.** Real-world gap, data quality/representativeness, interoperability.	assess beyond accuracy & cost, with lifecycle RWE and organisational/economic effects, early multi-stakeholder dialogue, broadened HTA scope. **Applied CEA hooks:** capacity/throughput, training/IT, incident/breach, over-testing/over-treatment.	None stated. To adapt an AI to a local environment, considerable investments and expenditures may be necessary. The evolution of AI in a real-world context of care and services, by integrating large amounts of data of various types and sources, requires additional resources to ensure its proper functioning and stability: continuous performance tests, software and data quality tests, infrastructure and equipment upgrades, human expertise, and training.

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Clinical effectiveness	Technical	**Lab results may not replicate in real-world**; few externally validated, context-fit studies ('AI chasm').	Prospective/externally validated studies in intended settings; pre-specified metrics & comparators.	Stepwise adoption; require Real world evidence confirmation pre-scale. Value of Information for more robust evidence; delay vs. harm avoided.
Safety and failure modes	Technical	**Black-box** errors can scale; drift to spurious signals.	Failure taxonomy; prompt sensitivity; adversarial/robustness tests; drift monitoring plan.	Post-market safety surveillance with triggers. Include in models adverse-event costs; monitoring program spend.
Data quality and generalisability	Technical	**Non-representative** data (sites/devices/populations) and leakage may lead to biased/fragile performance.	Dataset lineage; subgroup results; device/site transfer tests; leakage controls.	Data audits; re-evaluation on context change. Cost of reference test sets; misclassification burden.
Transparency and interpretability	Technical	Need to **explain**/justify outputs to support verification & trust; IP limits disclosure.	Model/method card (data, code/availability, rationale).	Document checks at appraisal and each update. Documentation effort

				vs. acceptance & error reduction.
Human factors (patients)	Procedural	Risks to **relationship, expectations, autonomy**; target population unclear.	Acceptability/usability evidence; informed-use labelling; target-population definition.	Feedback loops; complaint/override review. Include in model experience impacts; uptake support costs.
Human factors (clinicians)	Procedural	**Workflow burden, automation bias, cognitive overload**	Time–motion studies; override/recourse logs; training plans.	Phased rollout; proficiency checks; alert tuning cycles. Include in model productivity gains vs. training/overhead.
Organisational impact	Procedural	**Task altering, scope/jurisdiction** changes; pathway redesign.	Service redesign description; throughput and wait-time KPIs; staffing mix analysis.	Operational dashboards; trigger rules for pathway changes. Include in models capacity release; training and change-management costs.

Study: Ghabri (2025) : ‘Using AI in the Economic Evaluation of AI-Based Health Technologies’ (PharmacoEconomics, editorial)

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
Using AI in the Economic Evaluation of AI-Based Health Technologies (2025)	**Editorial / methods perspective** on: (i) what counts as **AI-based health tech** for HTA; (ii) how **AI can assist economic evaluation**; (iii) how to **report** AI economic evaluations (CHEERS-AI).	Treats **AI as enabling tech**; intervention is the **digital health tech embedding AI** (e.g., radiology analysis, DR screening apps).	AI decision support, screening /triage; also **GenAI assisting HTA tasks** (SLR, modelling).	**Fast iteration and learning curves, organisational change, dynamic pricing/versions**, equity/access issues.	Human-in-the-loop automation, Quality Assurance, bias/privacy/security controls, training. **CEA:** include costs of implementation/ maintenance, monitoring, productivity/time . Economic evaluation of AI needs to be **dynamic, transparent, and skills-aware,** with **CHEERS-AI** reporting and guarded use of	In economic evaluations of AI-based health technologies, distinguishing innovative health technology from the technological process (i.e., AI) facilitating the production of this technology is important. This distinction allows us to define innovative AI-based health technologies (e.g., digital medical technologies such as radiology image analysis to aid in diagnosing pneumothorax or eye disease screening apps for diabetic retinopathy, glaucoma, and macular degeneration) launched by manufacturers. These technologies are subject to HTA processes reflecting the requirements of each HTA agency for

					GenAI to assist HTA.	reimbursement and pricing decisions.

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Scoping & intervention definition	Procedural	**AI is not the same as the intervention (AI-based product);** with frontier AI the AI (and its **version**) is likely the assessed tech.	Clear intervention/comparator incl. **version/generation**; pathway role.	Re-scope when version or pathway changes. Include in the model re-certification/upgrade costs; displacement effects.
Dynamic change & learning curve	Technical	**Rapid updates** and **user learning** alter effectiveness/costs over time.	Time-varying performance and utilisation assumptions; learning-curve parameters.	Periodic re-calibration; trigger rules. Include the cost of update/maintenance; onboarding/training; downtime.
Evidence generation with GenAI	Technical	Large Language Models can aid **SLR/extraction,** but risk **hallucination**; must be augmentation only.	Declare GenAI use; prompts/logs; **accuracy vs. human** benchmarks.	Quality Assurance gates; replicate when models change. Time saved vs. Quality Assurance /rework.
Real world Evidence and parameterisation	Technical	AI aids **unstructured data extraction**; risk of misclassification/overfitting.	Subgroup accuracy; provenance; validation on hold-out sites/devices.	Data audits; bias/privacy reviews. Misclassification

				burden; security costs.
Model construction and uncertainty	Technical	GenAI can **draft structures/code**; **premature** to automate end-to-end modelling.	Side-by-side replication compared gold standards; structural/parameter uncertainty plan.	Pilot low-risk; escalate as pass rates improve. Debug/validation effort; Value Of Information for more data.
Reporting standards	Procedural	Need **CHEERS-AI** extension: **38 items** (28 CHEERS-2022 + 10 AI-specific).	CHEERS-AI checklist in appraisal; transparency on AI use/limits.	HTA compliance checks.
Equity and access	Procedural	Skills/affordability gaps may **limit equitable access.**	Distributional analysis; uptake by subgroup; affordability assumptions.	Equity checkpoints; mitigation plans.
Governance: privacy, security, ethics	Technical/Procedural	Economic evaluation must reflect **privacy/cybersecurity** and **ethical** safeguards.	Security protocol; disclosure of AI use; limits/warnings.	Regular audits; breach drills. Include in models breach expected loss; security program spend.

Study: Consolidated Health Economic Evaluation Reporting Standards for Interventions That Use Artificial Intelligence (CHEERS-AI) : Value in Health, 2024

Source (short	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype	Summary of key Frontier	Summary of key Frontier solutions	Example involving a case where AI technology faces

title; year)			(e.g., diagnostic assistant)	challenges (how HTA issues differ vs non-frontier)		a challenge relevant to frontier AI
CHEERS-AI reporting standard; 2024	Reporting **guideline** for economic evaluations of AI-based interventions (38 items: CHEERS-2022 + 8 elaborations + 10 new AI items)	No ‘frontier AI’ term. Usable mapping: items explicitly cover **locked vs adaptive (‘learning’) AI**, versioning, AI uncertainty, implementation:directly applicable to **frontier (LLM/LMM/FM) systems** with fast updates.	Any AI-based intervention (e.g., **LLM copilot**, **LMM diagnostic**, **AI triage/monitoring**) being appraised for cost-effectiveness	Frontier raises needs to **declare version/state, learning over time, AI-specific uncertainty, full cost map & implementation**, beyond typical non-frontier ML.	Frontier raises needs to **declare version/state, learning over time, AI-specific uncertainty, full cost map & implementation**, beyond typical non-frontier ML. **Applied CEA** (methods/inputs transparency for robust ICERs). Make CHEERS-AI **mandatory** for AI-HT EEs so HTA can trust inputs, trace versions, and price implementation & updates.	Not stated. Although some CHEERS-AI items are likely to be ‘future proofing’ the reporting standards against future developments of AI in healthcare, the authors recognize that this is a rapidly evolving field. We consider that major developments should be monitored and, if needed, CHEERS-AI may need to be amended or expanded over time.

MD: Medical Devices; US: United States; FDA: Food and Drug Administration; RWE: Real-world evidence

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Intervention description and comparators	Procedural	Must state **locked vs adaptive**, **version under evaluation**, user role/autonomy, intended purpose and workflow impact.	AI technique; locked/adaptive; version; intended users; added requirements; role in care; benefit rationale.	Bind decision to declared version/state; re-review on change. Include in costs in models implementation/setup; change-control overhead.
AI effect and learning over time	Technical	Frontier tools may **learn/drift** leading to effect sizes change post-launch.	AI effect data sources; **measurement and modelling of learning**.	Require pre/post evidence and learning assumptions audit; shorter refresh cycles. Re-validation; monitoring; rollback downtime.
Validation and bias	Technical	Larger generalisation/bias risk	AI **validation**, **population differences** train vs assess.	Equity-split reporting; site acceptance testing..
Uncertainty	Technical	**AI-specific uncertainty** (stochasticity, version drift) widens ICER ranges.	General uncertainty and **Impact of AI uncertainty**.	Require structural/scenario analyses for version/update pathways. PSA breadth; Value of Information for re-assessment cadence.
Costs and implementation	Procedural	Frontier adds **recurring** infrastructure, maintenance, retraining, licensing.	**Measurement/valuation of costs** elaboration: purchase price breakdown; implementation & maintenance.	Mandate full cost map and price evolution disclosure. Include in models cost of setup/training/IT; monitoring; licence/API; updates.

User autonomy and safety	Tech and Procedural	LLMs can be directive; **human-in-the-loop** must be explicit.	**User autonomy**; **Implementation** requirements and implications for CE.	Evidence of oversight points; incident reporting plan. Supervision workload; incident handling. Include in models cost of supervision workload; incident handling.

Study: Ethics and governance of AI for health: Guidance on large multi-modal models (LMMs) : WHO

Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
WHO Guidance on LMMs (2024)	WHO guidance on ethics and governance of **large multi-modal models** (aka 'general-purpose foundation models').	LMMs = a type of generative AI that accepts multiple input types and produces diverse outputs; often equated to **general-purpose foundation models.**	General-purpose LLM/LMMs across: clinical support, patient assistants, admin, education, research and drug development.	Identifies **systemic risks** more acute for LMMs: hallucination, bias, automation bias, skills degradation, privacy, affordability/access, labour impacts, cybersecurity,	**R**isk controls and duties by actor; audits, disclosure, procurement levers, participation, training. **Applied CEA hooks:** productivity, incident losses, privacy/security costs, environmental impacts.	It had been hoped that clinicians could use AI to integrate patient records during consultation, to identify at-risk patients and as an aid in difficult treatment decisions and to catch clinical errors (1). LMMs could make it possible to extend use of AI-based systems throughout diagnosis and clinical care – both virtual and in-person consultations, with some

				carbon/water footprint, epistemic authority shifts.	Stresses ethics, transparency, audits and lifecycle oversight for high-capability, fast-diffusing models.	experts expecting that LMMs ‘will be more important to doctors than the stethoscope in the past

MD: Medical Devices; US: United States; FDA: Food and Drug Administration; RWE: Real-world evidence

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What’s different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Clinical effectiveness and safety	Technical	**Hallucinations, bias, prompt-sensitivity** and data quality issues raise error risk in clinical and patient-facing use.	Task-specific accuracy with **subgroup** results; leakage controls; documentation of known failure modes; human-oversight plan.	Phased adoption; incident logging & **post-release audits**; trigger re-review with updates. Adverse-event/incident costs; oversight staffing; delay from re-testing vs. harm avoided.
Transparency, provenance and disclosure	Procedural	Training data and methods often opaque; outputs can look authoritative even when wrong.	‘Method/system card’ (training data classes, provenance, limits), **operational disclosures**, and labelling of AI-	Require disclosure at appraisal and on major change; procurement clauses for transparency. Documentation and audit program costs vs. acceptance and trust.

			generated content.	
Bias, equity and access	Technical/Procedural	AI can **amplify systemic bias** and widen digital divide; accessibility/affordability concerns.	Disaggregated performance; access and affordability metrics; disability-aware evaluations.	Equity checkpoints; public participation; corrective actions in deployment. Distributional CEA; subsidy/training costs; avoided inequity harms.
Human factors: automation bias and skills	Procedural	Risk of **automation bias** and **skills degradation** among clinicians; altered patient–clinician interaction.	Human-in-the-loop design; override/recourse logs; usability & acceptability studies.	Training and proficiency checks; monitor overrides and complaints. Training/uptake costs; productivity gains; quality-of-care externalities.
Privacy and data protection	Technical/Procedural	Patient data used via AI risk **leakage/misuse;** unclear retention and cross-use.	Data-protection impact assessments; clear terms for user-input data; privacy safeguards.	Periodic Data Protection audits; breach reporting and remediation plans. Security spend; expected loss from breaches.
Cybersecurity and information integrity	Technical	Larger **attack surface**; misinformation risks and deep content generation.	Threat model including prompt-/data-injection; content-integrity controls; resilience testing.	Continuous monitoring. Secuirty costs; downtime/incident losses.
Organisational impact and labour	Procedural	Workload reconfiguration, potential **job loss/retraining** needs;	Time–motion studies; role redesign and	Phased rollout; workforce engagement and feedback loops. Include productivity

			training plans; burnout metrics.	gains vs. retraining/change-management costs.
Environmental impact	Technical	**Carbon and water footprints** are material for training/inference.	Reported energy/water estimates; mitigation (model choice, efficiency).	Periodic sustainability review; vendor criteria. Energy/water costs in Budget Impact Analysis; carbon pricing where relevant.
Governance, liability and international rules	Procedural	**Developer/provider/deployer duties**, liability presumptions/strict liability options, and **international governance.**	Governance plan mapped to value chain; incident/impact assessments; evidence of regulatory compliance.	Public reporting; third-party audits; cooperate with international frameworks. Assurance and audit costs; liability/compensation funds.

Study: Aquino et al., 2024 : ‘Defining change: Exploring expert views about the regulatory challenges in adaptive artificial intelligence for healthcare’ (Health Policy & Technology, 2024).

Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI

Defining change: regulatory challenges in adaptive AI (Aquino et al., 2024)	Qualitative multi-stakeholder interview study (72 experts from Australia, Canada, NZ, US, UK)	Does not use ‘frontier AI’ term but studies adaptive / continuously learning AI, which matches the frontier AI category: Adaptive learning, cloud-based deployment, multi-component systems.	Adaptive clinical AI decision support / diagnostic assistant	Adaptive AI differs from locked algorithms by changing post-deployment without explicit human intervention; raises new issues around defining ‘significant change,’ responsibility (no clear definition of what counts as significant change or who monitors it.), and evidence persistence.	Criterion-based definitions (performance, risk, indication), shared responsibility, continuous monitoring. Applied CEA: may need to cost lifecycle governance and recertification. Adaptive AI blurs product boundaries and requires continuous, criteria-based regulation and governance to protect safety without blocking innovation.	These adaptive systems can change after deployment through interacting with new data, such that their responses are not always predictable. Studies have demonstrated the potential applications of adaptive algorithms in MRI-based breast cancer diagnosis, personalised glucose monitoring for diabetes management, and predicting clinical appointment delays

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What’s different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Health problem and current use	Procedural	Adaptive ML systems can change automatically without	Clear statement of intended use and population;	Periodic scope review when model behaviour or data context changes.

		explicit human intervention (i.e. without re-approval)	plan for change notifications when function increases.	Include in HTA models the costs of re-training, re-scoping, and clinical re-validation.
Description and technical characteristics	Technical	Multiple components (model, data, code, cloud infrastructure) create diffuse regulatory responsibility.	Component map with version control and data provenance audit.	Re-certify when major component is replaced or data pipeline changes. Include in HTA models the costs of lifecycle maintenance and compliance costs.
Safety	Technical	Behaviour can change after deployment with new risks or bias emerging over time.	Document known failure modes and post-market safety monitoring results.	Continuous real-world safety surveillance and incident review schedule. Include in HTA models the expected harm mitigation and safety-monitoring expenses.
Clinical effectiveness	Technical	Effectiveness evidence may expire as model learns from new data.	Evidence of ongoing performance validation against reference datasets.	Set triggers for re-evaluation when accuracy drops or data alters. Include in HTA models the costs of costs of re-testing and potential benefit loss from drift.
Ethical analysis / Legal aspects	Procedural	Lack of clarity about responsibility for monitoring change (clinicians vs developers vs regulator). Uncertainty over who is responsible when adaptive	Evidence of responsibility allocation and user/patient notification of AI adaptivity. Legal liability map and	Governance audit for accountability and consent compliance. Re-review legal accountability after major updates or jurisdiction changes. Include in HTA models the

		systems change (clinicians, developers, vendors, regulator).	data-protection impact assessment.	costs of ethical oversight and communication costs.
Organisational aspects	Procedural	Adaptive updates can disrupt clinical workflow and require new training.	Evidence of staff training and workflow adaptation plans.	Re-training schedule linked to software updates. Include in HTA models the costs of staff time, change-management and downtime costs.

Study: Toward the Autonomous AI Doctor (Hayat 2025)

Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI

				issues differ vs non-frontier)		
Toward the Autonomous AI Doctor (Hayat 2025)	Retrospective observational evaluation / benchmarking study using 500 real urgent-care telehealth encounters	No formal frontier definition, but system exhibits frontier-type attributes: fully autonomous clinical workflow, multi-agent orchestration (>100 agents), end-to-end reasoning	Autonomous clinical LLM performing full encounter (diagnostic and management and documentation)	Autonomy creates challenges not seen with narrow AI: evaluation depends on concordance rather than correctness, LLM-as-judge bias risk, anchoring bias when clinicians review AI notes first, documentation consistency may mask omissions, limited generalisability (single setting), no outcome validation	Study does not propose HTA solutions, but implicitly suggests: human adjudication for disagreements, structured error taxonomy, transparent reporting of agent roles, comparison against clinician documentation, and explicit noting of limitations (anchoring, no outcomes, generalisability)	Doctronic is a cloud-native modular system that has over 100 LLM-powered agents, each with a distinct, well-defined clinical role, mirroring the structured responsibilities of a human care team. These agents operate cooperatively, passing context-rich data to one another dynamically. The entire system is designed to mimic the clinical tasks of a primary care doctors office. Within this context, Doctronic is capable of performing a full medical history, after which it generates a SOAP note with the following components 1) a summary of the HPI including self reported physical findings and diagnostic tests, 2) a differential diagnosis with at least 4 diagnosis, 3) A plan for further diagnostic evaluation and treatment

						based on the differential. For the patient, the experience is similar to a two-way, open-ended, text-based chat.

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Intervention description & comparators	Procedural	First evaluation of a fully autonomous, multi-agent AI doctor performing complete clinical encounters (history, reasoning, plan, documentation). System includes >100 agents and mimics full primary-care workflow.	Full system description: agent roles, autonomy level, sequencing of tasks, Standard Operating Procedures for documentation, constraints applied to each agent.	Re-review when architectural modules or agent roles change; tie HTA to specific system version and agent configuration.
Evaluation method & adjudication approach	Procedural / Technical	Use of LLM-as-judge adjudication and human adjudication is novel; introduces potential systematic bias; LLM judge sometimes misclassifies clinically equivalent diagnoses when human wording is imprecise.	Transparent prompts, judge-validation evidence, correlation tests between LLM-judge and human adjudicators; error-taxonomy	Require independent human adjudication sampling; periodic calibration audits of LLM-judge performance; re-evaluation when judge model or prompts change.

			definitions; justification for equivalence criteria.	
Diagnostic and management agreement metrics	Technical	Frontier system evaluated on concordance not correctness. Plans judged for compatibility rather than accuracy; clinicians saw AI note first (anchoring risk).	Concordance tables, error analyses, proportion of discordances, expert-review summaries; explicit breakdown of ‘clinically equivalent but lexically different’ cases.	Mandate follow-up outcome studies; assess anchoring effects with alternative sequencing; schedule post-deployment validation using prospective outcome measures.
Safety (hallucinations, harmful errors)	Technical	Claims near-zero hallucinations and no harmful errors across 500 cases; evaluation depends on transcript consistency and LLM-judge framing.	Error-taxonomy definitions, count of hallucinations, examples of discordant cases, safety-related adjudication criteria.	Require independent safety audits with outcome follow-up; repeat safety evaluation when model, training data, or safeguards change.
Documentation quality & consistency	Technical / Procedural	Frontier system produces highly structured, consistent notes; very different surface wording vs clinicians but high semantic similarity. Raises	Textual similarity metrics, semantic similarity	Periodic audits for clinically relevant omissions; verification of how structured consistency

		question of whether consistency masks missing nuance.	metrics, human review of documentation adequacy; examples of omission cases.	affects clinical reasoning or billing workflows.
Generalisation & real-world representativeness	Technical	Evaluated only in U.S. urgent-care telehealth for adults; English-language only; limited generalisability acknowledged.	Case-mix description, demographic coverage, limitations statements, comparator clinician qualifications.	Require new HTA cycles for new languages, new settings (in-person visits), paediatrics, and high-acuity conditions.
Bias in evaluation & anchoring	Procedural	Clinicians were exposed to AI-generated notes before their own evaluation. This means possible anchoring bias inflating concordance.	Evidence on sequence effects; justification for design choice; human reviewer notes referencing anchor-related patterns.	Repeat evaluation with reversed order; require HTA sensitivity analysis on evaluator-sequence bias.
Outcome uncertainty	Technical / Procedural	Study explicitly does not measure correctness or patient outcomes only clinician agreement. True effectiveness remains unknown.	Explicit reporting of missing outcome data; justification for using concordance as	Require post-visit patient-outcome follow-up; define re-assessment cycle once patient-level outcomes are available.

			proxy; description of discordant-case review.	

Study: Oettl, F.C., Zsidai, B., Oeding, J.F. and Samuelsson, K., 2025. Artificial Intelligence and Musculoskeletal Surgical Applications. HSS Journal, 21(3), pp.267-273.

Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
Oettl 2025 – AI in Orthopedic Surgery	Clinical review article (not HTA-focused)	The article describes large language models, multimodal AI, agentic AI, continuously updating	Diagnostic assistants, intraoperative navigation systems, robotic or agentic	The article raises challenges that differ from conventional narrow AI, including rapid	The article implies solutions including sustained human oversight for autonomous or agentic functions, adoption of	One of the most promising orthopedic applications of agentic AI is its potential integration into robotic-assisted surgery. These systems could analyze real time intraoperative

		systems, intraoperative real-time models, and highly autonomous robotic-AI interactions. These features align with frontier AI characteristics such as rapid iteration, multimodality, environment-responsive autonomy, and broad system-level impact.	surgical automation modules, multimodal models for imaging and video analysis, and LLM-based clinical documentation and patient-interaction tools.	update cycles that undermine the durability of validation; real-time autonomous or semi-autonomous behaviours requiring strict oversight; multimodal and complex data dependencies that elevate issues of bias, generalisability, and system drift; hallucination and reliability limits in LLMs; and vendor dependence and infrastructural reliance for robotic and agentic systems. The article also	explainable AI to enhance transparency and safety, rigorous pre-use validation of imaging, navigation, and LLM-based systems, continuous monitoring of real-world performance through postoperative wearable-sensor and computer-vision data, ethical governance over autonomy and bias, and structured integration practices for robotics and multimodal systems.	feedback, imaging data, and surgical plans to enhance precision and adaptability during procedures. For example, an agentic AI could adjust robotic movements during a total joint replacement to account for unexpected anatomical variations, improving outcomes and minimizing surgical errors. Similarly, these systems could play a pivotal role in preoperative planning by synthesizing imaging and patientspecific data to create optimized surgical blueprints. Agentic AI is also poised to revolutionize postoperative care. Autonomous systems could monitor patient recovery using wearable sensors and imaging, providing early detection of complications such as infection or prosthetic loosening.

				highlights emergent behaviour risks in high-stakes surgical settings where black-box vision models or agentic systems influence intraoperative decision-making.		

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Description & technical characteristics	Technical	Multimodal, real-time, continuously updating systems (U-Net, YOLO, LLMs, agentic models, robotics). These exhibit higher autonomy, faster iteration, and opaque decision processes compared with traditional fixed-model ML.	Version-specific documentation of the model architecture, training data, update behaviour, interaction with robotic hardware, and real-time processing characteristics.	Re-validation after each substantive model, data, or autonomy update; requirement for updated technical dossiers reflecting multimodal inputs and real-time behaviours.

			Evidence of explainability approaches.	
Safety & incidents	Technical	The paper highlights silent failure risks in agentic or real-time models, hallucination in LLMs, and dependence on black-box detection models in intraoperative environments. These create elevated safety burdens compared with static diagnostic AI.	Stress-testing under worst-case or intraoperative edge cases; documentation of overrides; human-in-the-loop safety evidence; postoperative complication-detection performance.	Real-time incident monitoring and event taxonomies; pause/rollback triggers for agentic or robotic functions; periodic post-market safety review tied to evolving model behaviour.
Clinical effectiveness	Technical	Continuous data-dependence (vision, sensors, imaging) and evolving autonomy levels mean performance may alter across patients, settings, or updates.	Multi-site external validation including different imaging equipment, patient populations, and surgical settings; demonstration of consistency under real-time conditions;	Time-limited approvals with re-review following updates, changes in autonomy level, or detected drift; prospective monitoring of functional outcomes and complication detection.

			postoperative monitoring evidence from wearables and CV.	
Evidence generation (validation / RWE)	Technical	The article shows that AI systems used preoperatively, intraoperatively, and postoperatively evolve quickly and rely on multimodal data, making earlier validations rapidly stale.	Version-linked validation sets; external datasets for surgical-phase recognition, 3D planning, intraoperative navigation, postoperative gait analysis, and wearable-sensor outputs; evidence that LLMs are accurate and non-hallucinatory for documentation.	Scheduled re-tests or ongoing RWE pipelines; requirement to update validation when imaging/robot systems or LLM models change.
Equity, transparency & bias	Technical + Procedural	The article notes that LLMs and multimodal systems depend on large, heterogeneous datasets, raising concerns about bias in imaging, documentation, and patient-interaction tools.	Bias evaluation across populations; transparency documentation for LLMs and agentic models; evidence on	Mandated fairness/bias KPIs; periodic reviews of model outputs for demographic divergence; updating mitigation strategies as datasets or autonomy evolve.

			generalisability of imaging, surgical video, and postoperative sensor models.	
Interoperability & data integration	Procedural	The article emphasises strong dependence on imaging systems, surgical video feeds, robotics hardware, and wearable sensors, meaning integration failure poses operational and clinical risk.	Evidence of compatibility with imaging/video modalities, robotic platforms, and sensor ecosystems; cybersecurity and data-quality assurances for multimodal pipelines.	Site acceptance testing before rollout; staged deployment tied to infrastructure readiness; periodic reassessment when equipment or vendor systems change.
Organisational aspects & workflow	Procedural	Real-time AI models, agentic robotics, and LLM-based documentation reshape clinical roles, cognitive demands, and workflow patterns more extensively than narrow AI.	Workflow-impact evidence; impact analyses on skill needs, learning curves, documentation time, and dependency on robotic platforms.	Phased implementation with organisational readiness checkpoints; re-review if changes in autonomy or model updates alter workflow.

Costs & economic evaluation	Technical + Procedural	Continuous-learning surgical vision systems, postoperative monitoring AI, and agentic robotics introduce ongoing retraining, maintenance, and integration expenses beyond one-off capital equipment.	Lifecycle costing including compute, robotic integration, sensor management, maintenance of multimodal pipelines, oversight burdens, and costs of re-validation.	Dynamic cost-effectiveness models incorporating scheduled updates, monitoring, and infrastructure upkeep; re-assessment when autonomy levels or vendor systems change.
Ethical / patient & social considerations	Procedural	The article raises concerns about autonomy, explainability, LLM hallucination, and trust in agentic systems influencing clinical decisions and patient-facing communication.	Evidence of explainability for intraoperative and clinical-communication systems; demonstration that LLMs used in documentation or patient education meet accuracy and safety standards; consent and	Governance checkpoints for LLM and agentic-AI use; post-update re-review of patient-facing functions; monitoring of patient-trust and communication quality.

			transparency policies.	

Study: Panda, P.K. and Ghosh, S., 2025. Ethical use of AI in infectious diagnostic decision and therapeutic stewardship. *IDCases*, p.e02356

Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
Panda 2025 – Ethical challen	Commentary / ethics-focused scientific article	The article does **not** define 'frontier AI.' Instead, it	Diagnostic support systems; antimicrobia	Bias and fairness concerns; lack of	Human-in-the-loop oversight; transparent communication of	AI promises to transform diagnostics through enhanced accuracy,

ges and opportunities of AI in diagnostics and antimicrobial stewardship		discusses **LLMs, diagnostic AI, clinical decision-support models, autonomous treatment suggestions, and predictive models**, especially in TB, AMR, sepsis, oncology, and radiology. These are technically aligned with *frontier-adjacent* systems because they raise issues of bias, opacity, autonomy, and governance.	I stewardship decision-support; radiology models; LLM documentation tools	transparency; over-reliance on automated antimicrobial recommendations; accountability and liability; privacy and security risks; cultural appropriateness in LMIC settings; need for human oversight; risk of under-treatment/over-treatment driven by AI models; risk of misclassification in high-stakes settings; ethical risk from autonomous recommendations	AI limitations; bias audits; privacy safeguards; adherence to WHO, FDA, EU AI Act, and ICMR guidelines; context-specific validation; clinician training; community/ patient engagement; responsible governance	speed, and personalization. In radiology, AI systems have shown performance comparable to expert radiologists in detecting pneumonia on chest radiographs, lung nodules on CT scans, and breast cancers on mammography. In pathology, digital slide analysis using deep learning can identify malignancies with remarkable precision

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Bias, fairness & equity	Technical	AI models can replicate or worsen racial, demographic, or contextual biases (e.g., fracture risk misestimation). Bias is more consequential in general-purpose or widely deployed diagnostic AI.	Subgroup performance; demographic bias audits; context-specific validation for LMICs; evidence of mitigation strategies.	Periodic fairness audits; re-review if demographic drift is detected; require mitigation updates.
Human oversight & autonomy	Procedural	AI systems may influence or override clinical judgment, especially in antimicrobial stewardship or sepsis management. High-stakes decisions increase autonomy-related risk.	Documentation of human-in-the-loop workflows; evidence showing clinician oversight reduces errors; studies on override frequency.	Mandate human-in-command; monitor override patterns; re-review where automation bias appears.
Transparency & explainability	Technical	Many ML/LLM systems are opaque; clinicians cannot always understand the justification for antimicrobial or diagnostic recommendations.	Explanation or rationale suitable for clinicians; documentation of model logic, uncertainty and limitations; communication guidelines.	Update explanations with major model changes; evaluate clinician comprehension in periodic reviews.
Privacy & data protection	Procedural	Health and AMR/TB datasets are sensitive; large-scale	Data protection impact	Privacy audits; compliance checks aligned to national

		data use raises risks of breaches or misuse; contextual privacy concerns in LMIC settings.	assessment; consent model; privacy safeguards; cybersecurity documentation.	regulations (e.g., India ICMR guidance).
Accountability & liability	Procedural	Determining responsibility for misdiagnosis, inappropriate antibiotic recommendations, or sepsis treatment is difficult when AI is embedded in workflow.	Clear assignment of roles; evidence on human vs AI decision contribution; medico-legal mapping.	Governance checkpoints; role documentation; require re-review if autonomy level changes.
Cultural/context appropriateness	Procedural	AI trained in high-income settings may not generalise to LMIC contexts; risk of misalignment with local norms, resources, disease patterns.	Local validation studies; documentation of contextual adaptation; inclusion of local stakeholders.	Require local evidence prior to scale-up; periodic assessment of contextual fit.
Clinical risk & safety	Technical	Automated AMR recommendations or TB-severity assessments can cause patient harm if incorrect; risk of under-treatment.	Evidence of safety in real-world settings; pilot deployment data; error modes; risk communication plans.	Time-limited adoption with safety monitoring; re-review on performance drift; incident reporting.
Regulatory alignment & governance	Procedural	Panda emphasises WHO principles, FDA lifecycle AI, EU AI Act high-risk rules, and India’s ICMR guidelines. Systems must satisfy multiple regulatory regimes.	Demonstration of compliance with relevant guidelines; documentation of lifecycle risk management.	Governance gates; scheduled re-assessment aligned with regulatory updates; require documentation refresh.

Training & user preparedness	Procedural	Clinicians must interpret AI-driven recommendations and understand AMR/TB decision tools; risk of misuse without adequate training.	Training plans; competency evidence; usability studies; measures of comprehension.	Training refresh cycles; evaluate training effectiveness; link training completion to deployment.

Study: Wu, K., Wu, E., Rodolfa, K., Ho, D.E. and Zou, J., 2024. Regulating ai adaptation: An analysis of ai medical device updates. *arXiv preprint arXiv:2407.16900*.

Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA issues differ vs non-frontier)	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI
Wu 2024 – Regulating AI Adaptation	Preprint / policy-methods paper	Frontier-like AI = **adaptive / updating / post-deployment–changing medical AI systems**. Models that *learn or modify outputs after deployment*, can drift, respond to altering data distributions, or change via manufacturer	Risk prediction models, sepsis models, diagnostic imaging AI, adaptive classifiers; any AI whose performance varies across sites or time.	Evaluation is undermined because **models change after approval**; regulatory filings reflect outdated performance; **site-specific disparities**; unknown missingness; poorly	Continuous monitoring; enforce update logs; require post-market evaluation; mandate external validation across sites; implement methods for explainability and disparity analysis; governance for adaptive model changes; regulate real-world	The authors reference **sepsis prediction models** that performed substantially worse in external validations than in manufacturer appraisals, demonstrating drift and site-dependent failure after deployment. They also highlight **under-reporting biases** that change algorithm performance across patient groups (missingness problem).

		updates. Includes high-complexity clinical models with cross-site variability and distributional shift.		documented updates; lack of transparency in training data; distribution shift not captured by pre-market evidence; adaptive algorithms can create miscalibration and performance divergence; multi-site drift prevents static HTA validity.	performance tracking using claims and operational data.	

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Adaptive / changing model behaviour	Technical	AI models **change post-deployment** (developer updates, data-driven drift). Pre-market evaluation quickly becomes obsolete.	Version-specific performance metrics; documentation of each update; evidence on	Require **update-triggered re-review**; mandate continuous performance tracking and validation after each substantive model modification.

			how updates alter calibration and error rates.	
Cross-site variability and distribution shift	Technical	Algorithms behave inconsistently across hospitals due to differences in populations, clinical practice, equipment, and data distribution.	Multi-site external validation; subgroup and site-level performance reports; confounder SHAP analyses where applicable.	Scheduled **site-level monitoring**; automatic triggers for re-evaluation if site performance drops below thresholds.
Poor transparency of data, missingness, and under-reporting	Technical	Missingness is **unknown or unmeasured**, leading to biased performance and unexpected real-world failure.	Evidence describing missingness patterns; performance under different missing-data assumptions; robustness analyses.	Require routine reports of data quality and missingness; re-review when clinical data pipelines change.
Regulatory misalignment (static approvals vs adaptive AI)	Procedural	FDA/HTA processes assume 'locked' models; but frontier-like models update and drift, invalidating initial appraisals.	Post-market performance documentation; change logs; real-world surveillance data; insurance-claims–based	Move to **living HTA**; require periodic reassessment cycles; integrate RWE into ongoing regulatory evidence.

			adoption datasets.	
Lack of external validation and reproducibility	Technical	Manufacturers often submit limited or single-site evidence that does not generalize; external validations frequently contradict submitted results.	Mandatory external validation datasets; fully reported evaluation protocols; performance across demographic and clinical subgroups.	Require **pre-authorisation external validation**, then regular re-validation (every 6–12 months).
AI performance disparities & fairness issues	Technical	Models can produce **racial or site-based disparities** due to unobserved confounding or biased data.	Bias, fairness, and subgroup performance analyses; methods such as confounder-adjusted SHAP values.	Regular audits for bias drift; re-review if disparity metrics deteriorate.
Limited transparency of proprietary or black-box models	Procedural	Proprietary models block independent assessment of update mechanisms, training data, and performance assumptions.	Documentation of data sources, update mechanisms, and evaluation metrics; evidence explaining observed disparities.	Require transparency requirements; re-review when the developer changes model architecture, training data, or update cadence.

Inadequate real-world monitoring infrastructure	Procedural	Real-world performance is rarely tracked, but adaptive AI requires longitudinal monitoring using operational data.	RWE performance reports using claims, EHR data, operational metrics, and incident logs.	Implement continuous monitoring; establish thresholds for automatic regulatory intervention.

Study: Vielhauer et al., 2025 – Agentic AI System for pulmonary embolism (PE) work-up
Table A: summary of document

Source (short title; year)	Doc type	Frontier definition used/ aspects of frontier AI	Likely tech archetype (e.g., diagnostic assistant)	Summary of key Frontier challenges (how HTA	Summary of key Frontier solutions	Example involving a case where AI technology faces a challenge relevant to frontier AI

				issues differ vs non-frontier)		
Vielhauer et al., 2025 – Agentic AI System for pulmonary embolism (PE) work-up	Empirical validation study	No formal frontier definition, but the system uses large language models embedded in an agentic AI system with multiple autonomous steps, synthetic-data fine-tuning, and complex reasoning chains. These attributes represent frontier-type capability.	Clinical diagnostic decision-support agent operating within a multi-agent AI workflow	Frontier characteristics create challenges such as substantial performance variability across families and sizes, strong dependence on architecture and token limits, behavioural changes introduced by prompting and fine-tuning, error propagation across multi-step reasoning chains,	The study proposes systematic validation of each model-family and size, external benchmarking of tuning effects, structured tests for response consistency, synthetic-data pipelines to satisfy privacy requirements, and full end-to-end assessment of reasoning chains to identify where errors accumulate.	The decision support agent (DSA) was tasked to conduct an autonomous work-up of patients at risk of pulmonary embolism (PE). The DSA performance was tested against a large dataset from routine medical care, iterating the process using a cohort of open-weights LLMs.. The work-up of PE was chosen as the test scenario for two main reasons: (1) PE has significant clinical relevance and is underdiagnosed in routine medical care, and (2) our validation dataset contains sufficient decision-level clinical data on the work-up of patients with PE.

				unreliable therapy dosing recommendations, constraints on using real patient data for tuning, and significant compute and integration demands.		

Table B : Evaluation challenges and actions (Frontier vs non-Frontier)

HTA domain	Challenge type (Technical / Procedural)	What's different for frontier AI	Evidence required for HTA	Process action (monitoring/review schedule)
Model performance and diagnostic accuracy	Technical	Frontier models show marked variation in diagnostic accuracy across model families and parameter scales; accuracy depends strongly on architecture and token size, unlike fixed traditional algorithms.	Comparative accuracy across families and sizes, including accuracy, balanced accuracy, sensitivity and specificity as	Re-evaluation is required whenever the model family, size, or architecture changes because these changes produce new performance profiles.

			reported in the study.	
Tuning effects (prompting and fine-tuning)	Technical	Behaviour changes after fine-tuning; tuning improves some model families but reduces performance or reproducibility in others. This level of behavioural instability is characteristic of large frontier models.	Paired evaluations before and after tuning, error-type classification after tuning, justification of tuning procedures, and reproducibility testing.	Review is necessary after each tuning cycle; updates to prompts or fine-tuning data require new validation rounds tied to that specific version.
Response quality and reproducibility	Technical	Output quality varies by model; some produce incomplete responses, fail formatting requirements, or generate inconsistent answers across repeated runs. These features differ from deterministic non-frontier tools.	Response-quality scoring, reproducibility analysis using the study's identical-run tests, and documentation of formatting adherence.	Scheduling of reproducibility audits is necessary at each update, with logging of unstable or incomplete outputs.
Sequential decision-chain reliability	Technical / Procedural	The AI uses multi-step reasoning where earlier errors propagate to later stages; sensitivity decreases across steps because path-dependent mistakes accumulate. This behaviour is characteristic of agentic frontier architectures.	Step-wise performance metrics, chain-level balanced accuracy, and error-propagation analysis across workflow stages.	Review of full decision chains is required after significant model updates; pathway truncation patterns should trigger reassessment.
Therapy recommendation reliability	Technical	The models frequently recommend the correct medication but incorrect dose	Substance–dose–timing accuracy tables, error	Frequent re-evaluation of treatment outputs is required, and therapy-

		or timing. These high-stakes errors are a critical limitation of frontier LLM-based clinical systems.	classifications specifically for therapy recommendations, and comparison between families and sizes.	related functions may need restriction unless validated.
Data protection and training-data suitability	Procedural	Real clinical data cannot be used for tuning because of privacy constraints, so synthetic data are used. This creates uncertainty about whether synthetic-tuned performance transfers to real settings.	Evidence comparing synthetic-tuned performance with real-world validation results and documentation of the synthetic-data generation process.	Re-assessment of synthetic-to-real generalisation is required on a regular schedule, tied to changes in synthetic data pipelines.
Infrastructure and implementation burden	Organisational / Technical	Requires integration with clinical information systems and uses high-end compute infrastructure including large GPU clusters, which exceeds requirements for non-frontier clinical support systems.	System-latency data, implementation-resource mapping, compute-resource requirements, and integration performance characteristics.	Monitoring of infrastructure and integration requirements is needed; updates to the system or hardware require technical re-validation.
Model family selection	Procedural	Different LLM families exhibit distinct scaling behaviour, accuracy profiles, and tuning responses. Choosing a model	Family-level benchmarking data, documentation of	Reassessment at set intervals to ensure that model family selection

		family becomes a substantial design decision, unlike traditional algorithms.	architectural differences, and justification for selecting one family over alternatives.	remains appropriate as new LLM releases appear.