Medicines & Healthcare products
Regulatory Agency

Insights from the AI Airlock Simulation Workshops: Multistakeholder Perspectives on Explainability in AIaMD

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

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

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

Overview

Introduction

This is a report on a virtual workshop that we organised on 18th March 2025, as part of the AI Airlock pilot programme. The AI Airlock operates as a sandbox through which real world products can be explored collaboratively to develop a shared understanding of regulatory challenges posed by artificial intelligence (AI) technologies when used as medical devices, also known as AI as a Medical Device (AIaMD), and possible solutions to these.

AI systems, especially those based on neural networks or other complex models, often behave as "black boxes" whose internal logic is hidden from users. These models can make highly accurate predictions, but they do so through layers of computation that are difficult to interpret. In such "black box" models, we see inputs and outputs but cannot easily trace the logic and how an output was generated, making it difficult to verify the result.

AI has the potential for high utilisation in healthcare, including support for clinical decision makings. For example, AI-powered clinical decision support systems (CDSS) are emerging in radiology, pathology, and patient monitoring to help detect disease and recommend treatments. Even large language models (LLMs), pretrained on internet-scale data and often used as conversational assistants, are being tested for medical use. Recent work found that when prompted as a doctor, an advanced LLM could generate recommendations similar to those of a regulated medical device. Such examples suggest regulators and manufacturers may soon have to address explainability for these powerful new AI tools and explainability considerations might differ significantly depending on the type of AI model being used and the user. This need for clarity is especially important in high-pressure medical settings. In intensive-care units or emergency departments, clinicians report low confidence in recommendations from opaque AI systems. When decisions are life-critical, not understanding a tool's logic can lead to "decision paralysis" or misplaced trust.

Regulators recognise AI errors as a risk in healthcare. Given this context, the AI Airlock hosted a workshop titled "Explainability Requirements for Different Stakeholders." The meeting brought together experts from the NHS, government, academia, and industry to discuss challenges with explainability of LLMs and key questions, including what explainability means to different stakeholders, why it matters and the extent to which explanations should be provided.

A key outcome of the workshop was a set of recommendations for the MHRA as the medical device regulator, suggesting how official policy and guidance could be updated to ensure that AI-based medical devices consider the right level of transparency and explainability. This is critical for building trust among clinicians, patients, developers, payers, and other stakeholders, and for reducing the risk of harm to patients. These recommendations are interwoven in the discussion overview, but a list of recommendations for implementation are included in the AI Airlock Pilot Programme Report.

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

What Is "Explainability" in AI?

At its simplest, explainability means being able to state clearly how and why an AI system arrives at an output. In medical device regulatory guidance, standards such as BS/AAMI 34971:2023 (Application of BS EN ISO 14971 to machine learning in artificial intelligence) define explainability as the property of an AI system to present the important factors that influence a decision in a way that humans can understand. It introduces the need for adjusting explainability based on stakeholder needs, emphasising that algorithmic explainability should, where possible, include feature weightings. For example, a prediction model that takes in clinical measures such as BMI, cholesterol, blood pressure, should reveal how these 'features' are being used and what importance is assigned to them.

During the workshop, the group discussed whether different stakeholders should receive different types of explanations. The consensus was that tailored explainability is essential, even for developers as they may not fully understand black box systems.

Thus, this process needs to be understandable for different groups:

- **For developers:** Those who build AI systems need to understand how their tools work and how to fix any issues if things go wrong.
- **For regulators:** Regulators need proof that these systems are safe and work as they should. They need to know the details behind the decisions so they can ensure the technology meets strict safety, efficacy and quality standards.
- **For clinicians:** They need to be sure that the AI's suggestion fits with known medical practices. A clear explanation helps decide whether to trust the recommendation.
- **For patients:** Most patients are not interested in the technical details regarding AI decision making, but the meaningful, human-readable insight into how decisions affecting their health were made. This builds trust and ensures they can give informed consent to treatments.

The distinction between global and local explanations was discussed. Global explanations are useful for regulators and policymakers to offer insights into model behaviour across all predictions, while local explanations are more relevant to clinicians and end-users to clarify why a model made a specific decision for an individual case. This distinction supports explainability being context-dependent and role-specific. Workshop delegates suggested that in health contexts, over-explaining can be counterproductive, as information overload may reduce confidence rather than increase it. Conversely, insufficient detail may lead to blind trust or even misinterpretation of the AI's outputs. As such varying levels of technical detail are needed depending on the stakeholder. The group recommended that the MHRA should consider developing explainability criteria specific to different user groups. Striking the right balance of detail was identified as crucial, for instance, using interactive interfaces that allow users to "drill down" for more details only if they need it.

Explainability in Regulatory Guidance

Regulatory guidance such as BS/AAMI 34971:2023 (Application of BS EN ISO 14971 to machine learning in artificial intelligence) highlights that where full explainability is not possible due to the nature of an opaque system, justifications should be provided along with appropriate mitigations. In addition, the standard highlights the role of human involvement in decision-making, assessing different levels of human-in-the-loop interaction and their impact

on system outcomes. Good Machine Learning Practice (GMLP) Guidelines, particularly Principle 9, highlight the importance of providing clear and essential information to users based on their specific needs. These guidelines also emphasise the necessity of informing users about the scope and timing of device modifications and updates, as well as providing a way for users to communicate concerns to the manufacturer. Transparency for machine learning-enabled medical devices principles provide a structured approach by considering relevant audiences, motivations, the type of information that should be disclosed, its placement, timing, and the methods used to ensure transparency.

Why does explainability matter in healthcare?

Workshop delegates voiced that when all stakeholders, including clinicians, patients, developers, and regulators, can understand how an AI system reaches its conclusions, trust in the technology is more likely to develop. This is particularly important for clinicians who depend on AI to support clinical decision-making, and for patients who need assurance that their care is guided by transparent and well-reasoned processes. In addition, explainability enables the early identification of errors, inconsistencies, or biases within AI systems. By making it possible to interrogate how a decision was made, stakeholders can detect and address problems before they result in harm to patients. In clinical decision support, explainability allows clinicians to accept or override AI recommendations with confidence, especially critical in high-risk scenarios with potential for serious harm. For regulators, lacking insight into how algorithms function impairs the ability to predict failures and take appropriate action to protect patient safety.

More broadly, the group emphasised that explainability underpins fair, transparent decision-making, which is vital for public trust and procurement. Without that trust, the benefits of AI cannot be fully realised. Furthermore, transparent systems make it easier to determine responsibility when outcomes deviate from expectations. In alignment with this, user education was highlighted as a key pillar of AI use in healthcare, and explainability plays a pivotal role in supporting this. Earlier MHRA / CPRD workshops were referenced, which revealed clinician concerns around liability and litigation, with explainability being a key factor in building user confidence and adoption.

Challenges with explainability

Academic experts in the group highlighted the tension between desire for performance vs. demand for explainability. Explaining complex models is challenging, and perceptions around the need for explainability are highly context dependent. Academics have observed diverging viewpoints: some users are content to adopt AI systems based purely on performance, while others, especially those who believe in maintaining a human expert in the loop, demand explainability (why an output was predicted), interpretability (how the model is working), and clarity. Three key challenges with explainable AI (XAI) were identified from previous work conducted by the MHRA:

1. XAI methods often do not truly open the black box, instead adding an extra interpretive layer - this can introduce more complexity. The result can be a convoluted decision-making process, increasing the risk of human error.

2. Many AI-generated explanations either lack clarity or sufficient detail, making them ineffective for real understanding.
3. The idea that interpretability must come at the cost of accuracy is a myth - simpler, more transparent models may be just as effective and safer.

Explainability and humans in the loop

The group discussed whether AI explainability should vary based on device classification or the presence of a human in the loop. It was noted that clinicians may begin to doubt their own judgment when overriding AI outputs. One participant suggested that explainability needs are more dependent on use case than human involvement, while another highlighted the impact of AI literacy on user understanding. It was also argued that some, but limited, knowledge can be dangerous and make users overconfident. Systems should be designed assuming minimal user expertise to ensure safety.

The group discussed risks associated with over-reliance on AI-human collaboration, there is a need to design mechanisms that clarify how human-AI teaming works, rather than focusing solely on whether humans should be in the loop. Given the complex nature of some explainability approaches, tailoring it for different audiences is crucial as trust in technology is easily lost when systems are misunderstood. In addition, while human-in-the loop is often regarded as a suitable risk mitigation strategy, concerns were raised around how to practically detect errors in AI logic. Furthermore, how much effort should be dedicated to verifying that non-human “reasoning” is not flawed. Given this, it may be tempting to use AI to validate other AI systems, but some delegates warned that this could also present safety risks. One participant raised questions around this: if two AI models disagree, how do we know which one is right? What if both are wrong? And how can humans meaningfully intervene in a system too complex to oversee? This reinforced the need for explainability to be human-interpretable, not just machine-validated and further highlighted the need for upskilling alongside AI adoption.

Risks of insufficient explainability

Stakeholders reflected that insufficient explainability in AI medical devices can pose serious risks to patients. Concerns were raised that poor or absent explainability may lead to blind trust in AI, limiting users' ability to make informed decisions, particularly when data is incomplete or flawed. Similarly, when the system’s reasoning is not understood, there are also risks of loss of trust and user disengagement, particularly if outcomes differ from expectations. This makes it a relevant regulatory challenge where clarity will help users trust and use AI devices safely.

It was suggested that AI models should communicate both what they consider and what they omit through a model reasoning approach. The group re-enforced that poor explainability can lead to automation bias or defensive medicine, potentially resulting in harmful clinical decisions. Transparency tools, such as model cards – simple, structured overviews of how an advanced AI model was designed and evaluated – were recommended to help end users understand model limitations. In addition, explainability tools need to be carefully tested with end users to understand risks in more depth.

The current need for high levels of explainability was highlighted, driven by public trust concerns and the urgency to detect bias and model drift. However, it was suggested that as systems mature, the level of explainability needed at the patient level may decrease.

Different Ways to Explain AI Decisions through an AI Airlock case study: OncoFlow

A Real-World Example: The OncoFlow Tool

One example discussed during the workshop was the AI Airlock case study of OncoFlow, a system developed to support cancer treatment decision-making. OncoFlow integrates patient information with trusted medical guidelines and clinical data to generate treatment suggestions. It provides explanations through a combination of visual charts and clearly defined steps that show how each recommendation was reached. Additionally, the OncoFlow team developed a written explainability statement outlining how the model functions. This layered approach aims to support understanding among clinicians, regulators, and patients, each of whom may have different requirements and expectations.

The OncoFlow presentation outlined the system's two-model architecture: Model 1 handles data extraction, while Model 2 performs treatment matching. The workflow combines large language models (LLMs) and rule-based extraction techniques, drawing on sources such as NICE guidelines, clinical trials, formularies, and SACT datasets to identify appropriate treatments. The output includes a structured summary alongside proposed treatment options. The importance of data extraction accuracy was emphasised, and the availability of the explainability statement was noted as a step toward transparency and building trust.

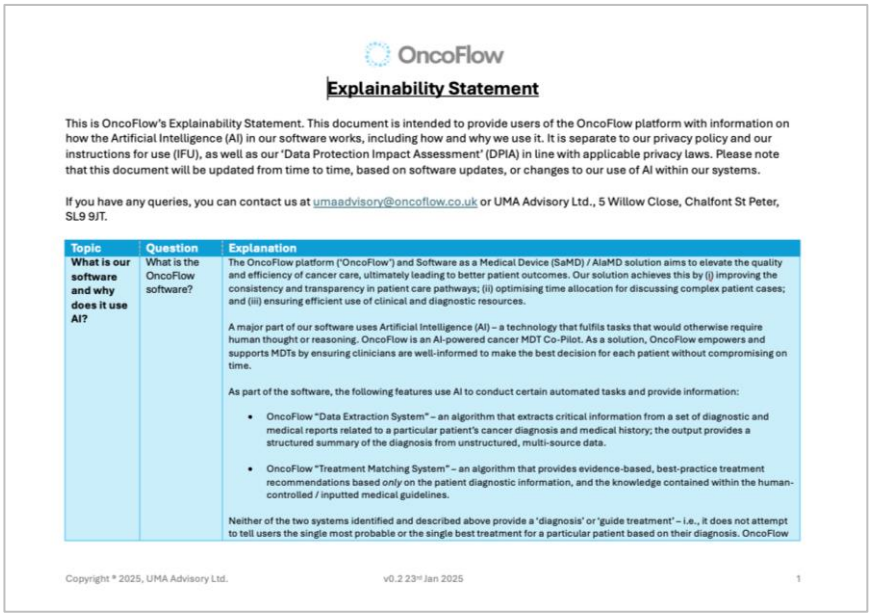
Using OncoFlow as a case study, delegates explored various explainability approaches in breakout sessions. Three main methods for making AI explainable were discussed, explainability statements, model reasoning, and quantitative approaches.

Explainability statements

Explainability statements are clear, written descriptions that summarise how an AI system makes decisions. They are designed to be easy to understand for people who are not experts. For instance, an explainability statement for a healthcare tool might explain that the system uses patient data and medical guidelines to suggest treatments. The key is to tailor the statement to different audiences, making it simple for patients but detailed enough for regulators and technical staff.

During the workshop, participants raised questions about what elements should be included in explainability statements. It was observed that current statements often lack sufficient detail about how users will interpret or interact with the information. There was a consensus on the need to tailor explanations to the needs of specific audiences, and to clarify what information is presented, how it is accessed, and by whom.

Participants suggested that a standardised or MHRA-endorsed template for explainability statements could help guide manufacturers. The Healthily's Explainability Statement was mentioned as a useful example. While acknowledging that product-specific differences will persist, participants agreed that a consistent structure would improve clarity and comparability.



OncoFlow

Explainability Statement

This is OncoFlow's Explainability Statement. This document is intended to provide users of the OncoFlow platform with information on how the Artificial Intelligence (AI) in our software works, including how and why we use it. It is separate to our privacy policy and our instructions for use (IFU), as well as our 'Data Protection Impact Assessment' (DPIA) in line with applicable privacy laws. Please note that this document will be updated from time to time, based on software updates, or changes to our use of AI within our systems.

If you have any queries, you can contact us at umaadvisory@oncoflow.co.uk or UMA Advisory Ltd., 5 Willow Close, Chalfont St Peter, SL9 9JT.

Topic	Question	Explanation
What is our software and why does it use AI?	What is the OncoFlow software?	The OncoFlow platform ('OncoFlow') and Software as a Medical Device (SaMD) / AIaMD solution aims to elevate the quality and efficiency of cancer care, ultimately leading to better patient outcomes. Our solution achieves this by (i) improving the consistency and transparency in patient care pathways; (ii) optimising time allocation for discussing complex patient cases; and (iii) ensuring efficient use of clinical and diagnostic resources. A major part of our software uses Artificial Intelligence (AI) – a technology that fulfils tasks that would otherwise require human thought or reasoning. OncoFlow is an AI-powered cancer MDT Co-Pilot. As a solution, OncoFlow empowers and supports MDTs by ensuring clinicians are well-informed to make the best decision for each patient without compromising on time. As part of the software, the following features use AI to conduct certain automated tasks and provide information: • OncoFlow "Data Extraction System" – an algorithm that extracts critical information from a set of diagnostic and medical reports related to a particular patient's cancer diagnosis and medical history; the output provides a structured summary of the diagnosis from unstructured, multi-source data. • OncoFlow "Treatment Matching System" – an algorithm that provides evidence-based, best-practice treatment recommendations based *only* on the patient diagnostic information, and the knowledge contained within the human-controlled / inputted medical guidelines. Neither of the two systems identified and described above provide a 'diagnosis' or 'guide treatment' – i.e., it does not attempt to tell users the single most probable or the single best treatment for a particular patient based on their diagnosis. OncoFlow

Copyright ® 2025, UMA Advisory Ltd. v0.2 23rd Jan 2025 1

Figure 2. Example of an explainability statement including breakdown and overview.

Model Reasoning

Showing how the AI thinks i.e. model reasoning involves breaking down the decision-making process into clear, visual steps. Think of it as a flowchart that shows the journey from input (like a patient's test results) to the final recommendation. This visual guide helps end users (clinicians in the case of OncoFlow) see the logic behind a suggestion and it allows them to spot any steps that might need further review.

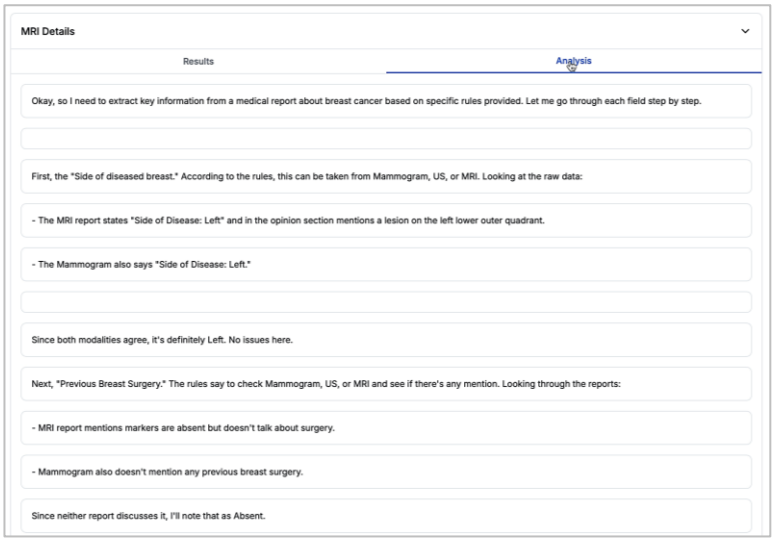


Figure 3. Example of model reasoning i.e., a breakdown of the logic that led to the model response.

In a discussion which focused on the role and value of model reasoning in clinical decision-making, participants acknowledged that while detailed reasoning may be time-consuming to review during initial use, it becomes less burdensome as teams grow more familiar with the tool. In both routine and complex cases, model reasoning was seen as a useful starting

point, particularly when paired with mechanisms to learn from outlier or edge cases. Several participants stressed the importance of this transparency around how information is sourced, even if the reasoning is not overly detailed. There was agreement that omissions in data sources must be made clear in the reasoning, so clinicians can assess the completeness of the information before acting on recommendations. Concerns were raised about the absence of patient-specific factors when a model extracts only information from a patient health record, information which is typically accounted for through clinician-patient relationships, e.g. social context, or personal preferences, may be missed. These human elements remain essential to MDT discussions and are difficult to encode into an AI system. Some participants proposed the inclusion of patient-facing tools - such as a diary - to help surface patient perspectives in MDT discussions where AI is used. When relaying information to patients, it was acknowledged that while patients generally do not require technical detail, they value clear, accessible explanations of their treatment options and potential side effects.

From a regulatory standpoint, participants noted that model reasoning should align with the device's intended use, associated controls, and regulatory requirements. Post-market surveillance (PMS) and ongoing performance monitoring were identified as important mechanisms to ensure long-term accountability. The potential utility of quantitative explainability techniques was discussed, with participants considering whether diverse methods could converge on coherent and trustworthy explanations. In terms of design, there was strong support for user interfaces that offer summary views with the option to "drill down" into further detail. As time is often limited, particularly in clinical settings, the ability to access more granular information when needed is critical. Ultimately, model reasoning contributes to transparency and trust, both of which underpin sound clinical decision-making.

Quantitative Approaches

These methods use numbers and charts to show which pieces of information were most important in making a decision. Techniques like SHapley Additive exPlanations (SHAP) can help developers check that the AI is working correctly and give regulators confidence that decisions are based on real data rather than guesses. However, the challenge is to balance these technical details with the need to keep explanations clear and simple for non-technical users.

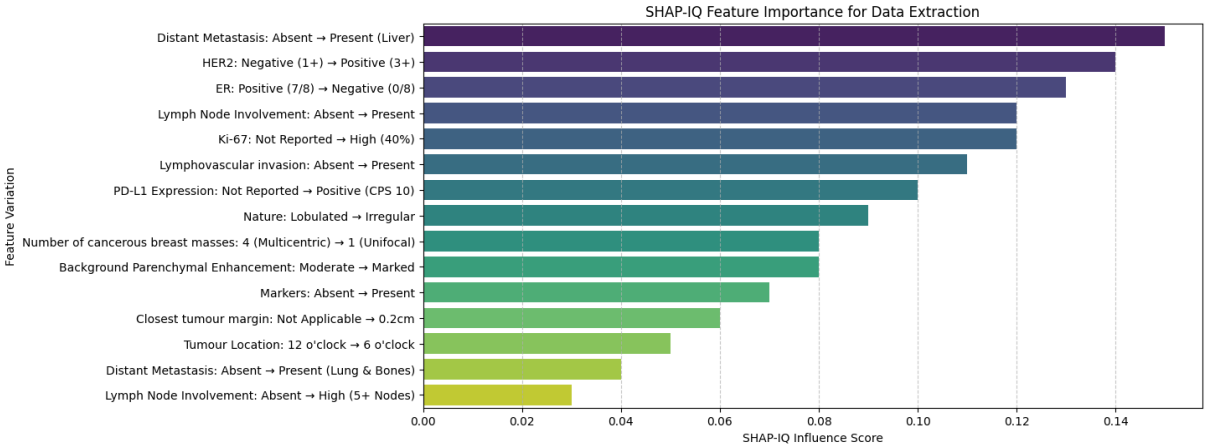


Figure 4. Example of quantitative approaches: SHAP bar chart illustrates how changes in the wording of clinical reports affect the medical information extracted by the AI. The graph shows which terms have the greatest impact on the output of the AI.

In discussion, it was suggested that a summary of all the approaches used across the cases would be helpful in narrowing down the most appropriate methods. Understanding the rationale for selecting each method, and its intended purpose, would provide essential context.

The discussion highlighted concerns about the reliability and complexity of quantitative model-based approaches, particularly in relation to clinical trust. Participants stressed the importance of understanding how results are generated and how models consider the complexity and nuance inherent in healthcare delivery. A key challenge identified was the potential mismatch between statistically derived factors and clinical reasoning. There was also discussion on how best to assess whether a quantitative methodology is suitable for evaluating explainability. Participants noted that quantitative methods such as SHAP and Local Interpretable Model-agnostic Explanations (LIME) themselves rely on underlying models that require their own explanations. As a result, interpretability must extend not only to the outputs but also to the methods used to generate them.

As AI becomes more widely used in healthcare, it was suggested that clinicians may grow more comfortable with various explainability tools. Visualisations, such as colour-coded outputs, summary charts, and other graphical representations, were seen as valuable for rapid interpretation. Participants emphasised the benefit of comparing outputs across methods to detect where they align or diverge, as this may signal consistency and reliability. An added benefit would be to assess the alignment between explainability outputs and what clinicians consider plausible based on experience.

Finally, participants distinguished between understanding the reasoning behind an individual output and understanding how the model itself functions. Ideally, both should be presented in a human-interpretable manner. Interest was also expressed in how user interface / user experience design could help ensure users are aware when guidelines or reference information have been updated in real time - a critical feature for maintaining clinical trust in the system's currency and accuracy.

Summary

This workshop highlighted several critical issues around explainability in AIaMD, including its importance, how it may need to vary for different stakeholders, and potential approaches for explaining AI outputs. The need to embed explainability from the outset of development for all users, both technical (e.g. developers) and non-technical (e.g. patients) was emphasised. The risks of insufficient explainability were noted as significant; it should not be treated as a compliance exercise or added on later. The ongoing tension between performance and explainability was acknowledged, and the context of use for the AIaMD remains a key factor in selecting appropriate explainability methods. Written statements, intuitive visual aids, and data-driven explanations to make AI systems understandable, safe, and accountable should be considered. As the role of AIaMD expands, a sustained commitment to clear and context-appropriate explanations will be essential to ensure the technology delivers meaningful benefits for all.

The workshop generated practical, evidence-based insights and recommendations, captured in the AI Airlock Pilot Programme Report. These will inform the future work of the National AI Commission and shape MHRA's regulatory strategy in Great Britain, while also contributing to international regulatory discussions.