

Understanding healthcare workers' confidence in AI

Report 1 of 2

May 2022

NHS AI Lab & Health Education England



This report has been developed by Health Education England and the NHS AI Lab at the NHS Transformation Directorate.

Health Education England

Health Education England (HEE) exists for one reason only: to support the delivery of excellent healthcare and health improvement to the patients and public of England by ensuring that the workforce of today and tomorrow has the right numbers, skills, values and behaviours, at the right time and in the right place.

At any one time, HEE supports more than 160,000 students and trainees whilst working closely with partners across the NHS locally, regionally, and nationally on shared priorities.

In 2019, HEE was commissioned by the then Secretary of State to deliver the Topol Review recommendations looking at the impact of leading-edge digital technologies on the workforce. The Digital, Artificial Intelligence and Robotics Technologies in Education (DART-Ed) programme picks up from this in 2021 to explore the linkage between mature evidenced AI and its workforce impact and required training and education.

NHS AI Lab

The NHS AI Lab was set up to accelerate the safe, ethical and effective adoption of AI in health and social care. Its vision is that the UK will be world-leading for the development and use of AI-driven technologies to improve people's health and wellbeing, delivering the most impactful technology to support our health and care system. The Lab creates an environment for collaboration and co-creation by bringing together programmes that address the barriers to developing and deploying AI in health and care. This will unlock the potential of AI to change the way healthcare is delivered, whilst ensuring we can determine the right guidance and regulations to protect patients and those in care.

About the authors

Dr Mike Nix is a Clinical Fellow for AI and Workforce at Health Education England and the NHS AI Lab.

George Onisiforou is a Research Manager for the AI Ethics Initiative at the NHS AI Lab.

Dr Annabelle Painter is a Clinical Fellow for AI and Workforce at Health Education England and the NHS AI Lab.

The authors would like to thank:

- » Dr Hatim Abdulhussein (Health Education England), Brhmie Balaram (NHS AI Lab), Lucy Dodkin (Health Education England), Tom Hardie (The Health Foundation), and Richard Turnbull (Health Education England) for their ongoing guidance and oversight of all research deliverables.
- » Dominic Cushnan, Louise Evans, Sarah-Jane Green, Eleonora Harwich, Alison Lowe, Giuseppe Sollazzo, Leanne Summers, and Mathew Watt (at the NHS AI Lab); and Prof Adrian Brooke, Alan Davies, Patrick Mitchell, and Chris Munsch (at Health Education England) for reviewing aspects of this report.
- » Allison Gardner (NICE), Dr Xiaoxuan Liu, Dr Russell Pearson (MHRA), Sir David Spiegelhalter FRS OBE, and Harriet Unsworth (NICE) for providing valuable feedback on aspects of this report.
- » the individuals listed in Appendix A who contributed to this research through interviews and feedback on this report (and have agreed to be acknowledged), and all other interviewees for this research.
- » Jennifer Berger (NHS AI Lab) and Beth Johnson (Health Education England) for their support with communications activities.
- » Nic Hinton (Karoshikula) for designing the report.
- » Hena Aziz (NHS AI Lab) for her support with research tasks.

Contents

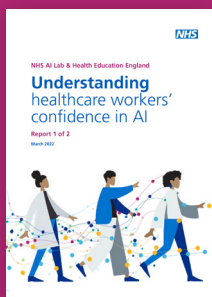
Executive Summary	5
Report overview	15
Chapter 1: Introduction	16
1.1 Research purpose	16
1.2 Methodology	17
1.3 Research reports	17
1.4 Context	18
1.5 Terminology	20
Chapter 2: Key Concepts and Framework	21
2.1 Key concept: Confidence in AI	21
2.2 Key framework: Understanding confidence in AI	23
2.3 The importance of developing confidence in AI among healthcare workers ...	27
2.4 The importance of developing confidence in AI among patients and the public	29
Chapter 3: Governance	30
3.1 Regulation and standards	31
3.2 Evaluation and validation	36
3.3 Guidelines	41
3.4 Liability	45
Chapter 4: Implementation	47
4.1 Strategy and culture	48
4.2 Technical implementation	51
4.3 Local validation	53
4.4 Systems impact	55
Chapter 5: Clinical Use	57
5.1 Aspects of clinical reasoning and decision making (CRDM)	58
5.2 Factors affecting confidence in AI during CRDM	61
5.3 Cognitive biases and appropriate confidence in AI-assisted CRDM	74
5.4 Interface with patients	80
Conclusion: Developing Healthcare Workers' Confidence in AI	82
Appendix A: List of interviewees	85
References	86

Executive Summary

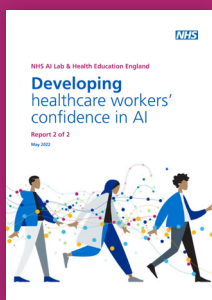
This research, which comprises two reports, is a collaboration between the NHS AI Lab and Health Education England. Its primary aim is to explore the factors influencing healthcare workers' confidence in artificial intelligence (AI) technologies and how these can inform the development of related education and training.

The research follows the Topol Review (2019) recommendation to develop a healthcare workforce able and willing to use AI and robotics, and is part of Health Education England's Digital, AI and Robotics Technologies in Education (DART-Ed) programme to understand the impact of advances of these technologies on education and training requirements. Supporting healthcare workers to feel confident in identifying when and how to use AI is a main objective of the NHS AI Lab, and a key component of its vision for the safe, effective, and ethical adoption of AI technologies across health and care.

This is the first of two reports in relation to this research.



This first report outlines a conceptual framework for understanding what influences confidence in AI among healthcare workers.



The second report will determine educational and training needs based on the findings and conceptual framework of this report, and present pathways to develop related education and training offerings.

The research involved a review of academic literature and semi-structured interviews exploring experiences of developing and using AI technologies in healthcare settings. Interviewees included healthcare workers in primary and hospital care settings; industry innovators; representatives of related regulatory and arm's length bodies; and academics who work at the intersection of AI, healthcare, education and clinical confidence.

AI in health and care settings

'AI' or 'AI technologies' describes the use of digital technologies to create systems capable of performing tasks commonly thought to require human intelligence. These can include algorithms using statistical techniques that find patterns in large amounts of data, or perform repetitive cognitive tasks using data, without the need for constant human oversight.

This definition of AI is intentionally broad and could encompass algorithms not commonly considered as AI. Many of the factors that influence confidence in AI discussed in this report could apply to any data-driven technology or algorithm used in healthcare or clinical practice.

AI technologies have the potential to support existing clinical capabilities in diagnosis and screening, drug discovery, digital epidemiology, and personalised medicine,¹ as well as optimising organisational resources, improving system efficiencies and clinical workflow pathways.

Most healthcare workers lack direct experience with AI technologies. A 2020 survey of over 1,000 NHS staff by the Health Foundation found that three-quarters of respondents have heard, seen or read 'not very much' or 'nothing at all' about automation and AI.²

However, the uptake and use of healthcare AI technologies is accelerating. An increasing number of technologies are expected to be deployed within the next three years. This is highlighted in Health Education England's survey of 240 AI technologies, where over 20 per cent of these technologies were estimated to be ready for large scale deployment within 2022, and an additional 40 per cent within three years.³

Moving from trust to appropriate confidence

The literature review and analysis of the interviews conducted for this research suggest that trust and confidence are often used interchangeably, with increased trust in AI often stated as a desirable objective in health and care settings.

Therefore, when considering how AI technology is used in healthcare, it is important to differentiate between the terms trust (which is placed in a product or system), trustworthiness (which is earned), and confidence (which is held individually or collectively). These distinctions have informed this report's conceptual framework, which uses the term confidence rather than trust.

The term **trust** is a belief in the reliability of a product or system and is typically a binary concept such that something is either trusted or it is not.

In this context, **trustworthiness** encompasses the quality of a product or system, being deserving of trust or confidence.

Confidence, like trust, conveys a belief in a product or system. However, unlike trust, it is not generally considered a binary concept. Instead, confidence can be understood as continuously variable and depending on various factors. Confidence can account for the nuances of using AI in clinical decision making, where high confidence in AI-derived information is not always a desirable objective. It allows for a more dynamic exploration of related influences and behaviours, including where lower confidence may be justified.

Interviews for this research suggest that confidence in any AI technology or system used in health and care can be increased by establishing its trustworthiness. Increasing confidence in this way is desirable and requires a multifaceted approach including regulatory oversight, real-world evidence generation and robust implementation.⁴

In the context of clinical decision making, once trustworthiness in AI technologies has been established, high confidence in AI-derived information (the output provided by an AI system to a clinician) may not always be desirable. Instead, different levels of confidence may be held in individual outputs from a given AI technology, depending on the context and circumstances. During clinical decision making, confidence in AI-derived information will depend on numerous factors including the clinical scenario and other available sources of information. The challenge, therefore, is to enable users to make context-dependent value judgements and continuously ascertain the **appropriate level of confidence in AI-derived information**, balancing AI-derived information against conventional clinical information.

A framework for understanding confidence in AI

This report presents a framework for understanding what influences confidence in AI within health and care settings, which was developed following analysis of the academic literature and the interviews conducted for this research.

Establishing confidence in AI can be conceptualised as:

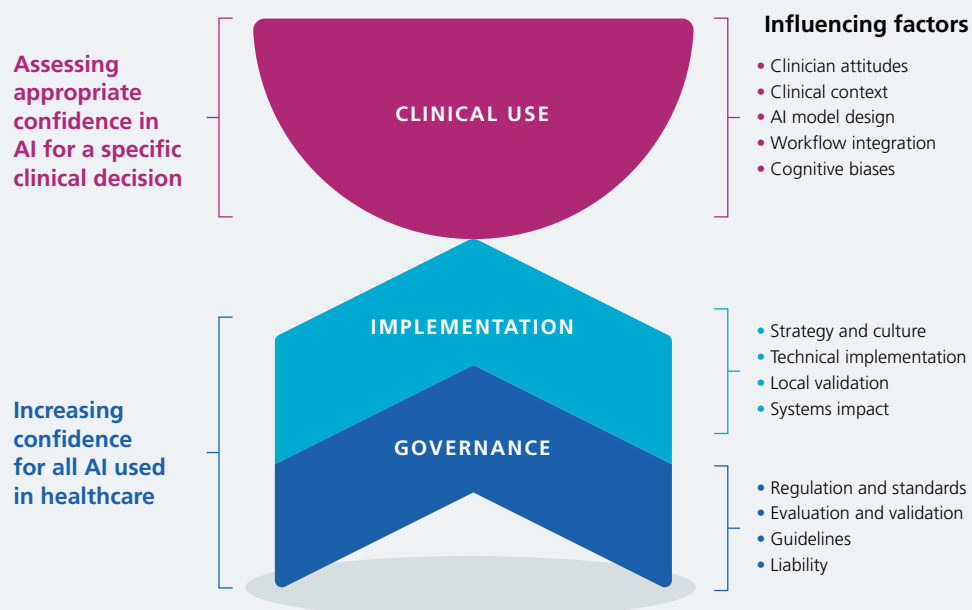
- » **Increasing confidence in AI by establishing its trustworthiness (applies to all AI used in healthcare)**
 - › Trustworthiness can be established through the **governance** of AI technologies, which conveys adherence to standards and best practice, and suggests readiness for implementation.
 - › Trustworthiness can also be established through the robust evaluation and **implementation** of AI technologies in health and care settings.
 - › Increasing confidence is desirable in this context.

» Assessing appropriate confidence in AI (applies only to AI used for clinical decision making)

- › During clinical decision making, clinicians should determine appropriate confidence in AI-derived information and balance this with other sources of clinical information.
- › Appropriate confidence in AI-derived information will vary depending on the technology and the clinical context.
- › High confidence is not always desirable in this context. For example, it may be entirely reasonable to consider a specific AI technology as trustworthy, but for the appropriate confidence in a particular prediction from that technology to be low because it contradicts strong clinical evidence or because the AI is being used in an unusual clinical situation. The challenge is to enable users to make context-dependent value judgements and continuously ascertain the **appropriate level of confidence in AI-derived information**.

Figure A illustrates the conceptual framework and lists corresponding factors that influence confidence in AI. These comprise factors that relate to governance and implementation, which can establish a system's trustworthiness and increase confidence. Clinical use factors affect the assessment of confidence during clinical decision making on a case-by-case basis.

Figure A: Framework for understanding confidence in AI among the healthcare workforce



The primary focus of this report is understanding and assessing appropriate levels of confidence in AI-derived information during clinical use. However, as appropriate confidence in AI for clinical decision making is premised on establishing the trustworthiness of these technologies, the key elements of governance and implementation that underpin confidence are addressed first before clinical use is discussed in more detail.

Governance

Increasing confidence through the governance of AI technologies

Robust governance underpins the trustworthiness of AI technologies, and can increase confidence in workers who commission, implement and use AI for any task in health and care settings. Aspects of such governance can include:

- » **Regulatory frameworks and standards.** Regulation and standards can provide assurance that AI technologies have been developed responsibly, work as advertised and are safe for users and patients. Interviewees for this research highlighted several areas where developments in regulation could increase confidence in AI. These include the regulation of AI technologies (through AI-specific medical device regulation), the regulation of healthcare settings (through guidance on the safe and effective use of AI technologies), and the regulation of professionals who develop, validate and use AI (through advice from regulators of healthcare workers).
- » **Evaluation and validation.** AI technologies classed as medical devices require internal validation for MHRA approval, but external validation, prospective clinical studies and, in some cases, local validation can build confidence in an AI's performance. Several standards and tools have or are being developed for medical devices and clinical research to guide approaches to the evaluation of AI products, including the National Institute for Health and Care Excellence (NICE) evidence standards framework.
- » **Guidelines.** Guidelines on the procurement, development and use of AI can enhance confidence when adopting AI in healthcare settings. Clinical guidelines from entities such as NICE and the Royal Colleges can also support confidence in using AI technologies.
- » **Liability.** Clarity on liability across different AI technologies is crucial to securing the workforce's confidence in using AI technologies. Currently, there is uncertainty as to who will be held to account if AI products are used to make clinical decisions that lead to patient harm. Responsibility could fall to the clinician who uses the technology, the deploying organisation, the industry innovator that developed the technology or those who validated and approved the technology for clinical use. Various legal frameworks may be applicable including negligence, product liability and vicarious liability. The NHS AI Lab's Regulations programme is exploring these issues in greater depth, including through its 'Liability and Accountability' portfolio of work.

Implementation

Increasing confidence through the robust implementation of AI technologies

Interviewees for this research highlighted that the safe, effective, and ethical implementation of AI in health and care settings underpins the trustworthiness of AI technologies, and contributes to the workforce's confidence in these technologies. Such implementation can include:

- » **Strategy and culture.** The leadership, management and governance bodies within health and care settings establishing AI as a strategic asset, and maintaining organisational cultures conducive to innovation, collaboration, and public engagement. This can include co-developing AI technologies 'from the ground up' with industry innovators, by engaging and involving multi-disciplinary teams (including clinicians, information technology and governance specialists, clinical domain experts and data scientists) and internal decision-makers early in discussions about their needs and implementation challenges.
- » **Technical implementation.** Addressing any technical implementation challenges involving information technology infrastructures, interoperability, and data governance requirements is also crucial. Interviewees noted that agreed information technology and governance arrangements are instrumental to healthcare workers' confidence in using AI technologies. A focus on the digitalisation of health and care services is an important prerequisite to the adoption of AI technologies, as supported by the NHS's What Good Looks Like framework.
- » **Local validation.** Procurement or commissioning entities within health and care settings will need to decide whether to validate the performance of AI technologies to ensure its performance translates to local data, patient populations and clinical scenarios. There are many unknowns and potential risks involved in 'translating' AI technologies from controlled development and validation settings to complex and highly individual real-world settings. These risks can relate to the ability of settings to understand the suitability and performance of the AI technologies locally, to maintain the ongoing rigour of that performance, and to minimise any unfair impact on, or harm to, patients.
- » **Systems impact.** Healthcare workers will be more confident in AI technologies that are safely and efficiently integrated into existing workflow systems that should include pathways for reporting safety events. An ethical approach to AI will also be essential to achieving confidence in AI. Interviewees noted that, at a minimum, this can include the principles of fairness, transparency, and accountability and ensuring equitable benefits across patients.

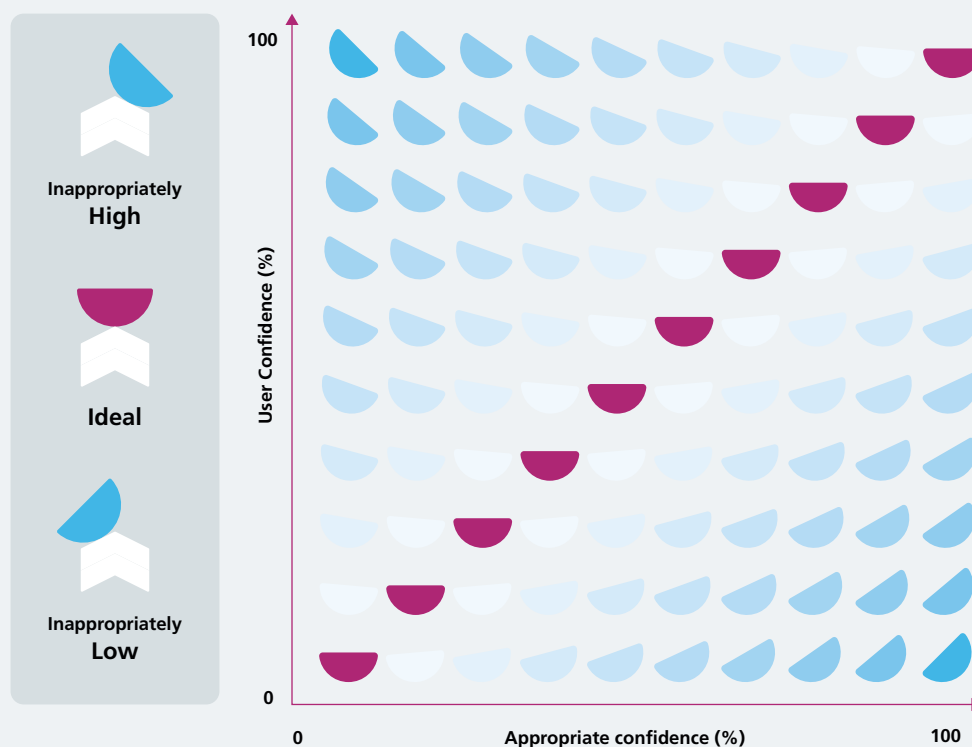
Clinical use

Assessing appropriate confidence in AI-derived information during clinical decision making

Figure B shows that a clinician's actual confidence (shown as 'User confidence') in AI-derived information for decision making may be inappropriately high or low if it does not match (along the 'ideal' line) the appropriate level for a clinical case. As discussed in this section, this appropriate level is likely to vary from case to case, depending on clinical factors and the AI-derived information itself.

To avoid inappropriately high or low levels of confidence, clinicians need to determine the appropriate level of confidence in the specific AI-derived information available at the point of making each clinical decision.

Figure B: Ideal and inappropriate levels of confidence



A complex set of considerations can dictate how to determine an appropriate level of confidence in AI-derived information, depending both on the technology and the clinical scenario, and with certain AI technologies and use-cases presenting lower clinical or organisational risks.

Synthesising and evaluating information from many disparate sources are key skills in clinical decision making, whether for diagnosis, prognostication or treatment. If incorporated and considered with appropriate confidence, information from AI technologies has the potential to make clinical decision making safer, more effective, and more efficient.

To achieve this, clinicians will need to understand when AI-derived information should and should not be relied upon, and how to modify their decision making process to accommodate and best utilise this information. This might include considering factors like:

- » other sources of clinical information and how to balance these with AI-derived information in decision making
- » the clinical case for which the AI is being used
- » the intended use of the AI technology.

Several factors can influence how clinicians view AI-derived information (their user confidence), potentially leading to inappropriately high or low levels of confidence. These include:

- » **Clinicians' attitudes.** General digital literacy, familiarity with technologies and computer systems in the workplace, and past experiences with AI or other innovations can influence assessments of confidence in AI-derived information.
- » **Clinical context.** The clinical context in which AI is used can influence confidence in the technologies and in the derived information, including in relation to the levels of clinical risk and the degree of human oversight in the AI decision-making workflow.
- » **AI model design.** Various design characteristics can influence confidence in AI technologies. For example, the way AI predictions are presented (such as diagnoses, risk scores, or stratification recommendations) can affect how clinicians process information and potentially influence their ability to establish appropriate confidence in AI-derived information.
- » **Cognitive biases.** Cognitive biases, including automation bias, aversion bias, alert fatigue, confirmation bias and rejection bias can affect AI-assisted decision making. The propensity towards these biases may be affected by choices made about the point of integration of AI information into the decision making workflow, or the way such information is presented. Interviewees for this research highlighted that enabling clinicians to recognise their inherent biases, and understand how these affect their use of AI-derived information should be a key focus of related training and education. Failure to do so may lead to unnecessary clinical risk or the diminished patient benefit from AI technologies in healthcare.

Developing healthcare workers' confidence in AI

Interviewees for this research stressed the importance of the healthcare workforce being confident in their own ability to adopt and use AI technologies.

Low confidence may limit the use of AI technologies and result in wasted resources, workflow inefficiencies, substandard patient care and potential disparities in who gets to benefit from AI technologies, which may be unethical.

During clinical decision making, inappropriate levels of confidence in AI-derived information could lead to clinical errors or harm in scenarios where the AI underperforms, without being properly assessed or checked. This includes a phenomenon known as automation bias where the user inappropriately and uncritically favours suggestions made by automated decision making systems.

The main recommendation of this report is therefore to **develop and deploy educational pathways and materials for healthcare professionals at all career points and in all roles, to equip the workforce to confidently evaluate, adopt and use AI**. During clinical decision making, this would enable clinicians to determine appropriate confidence in AI-derived information and balance this with other sources of clinical information.

The factors influencing confidence in AI, as detailed in this report, can help to determine the educational requirements to develop such confidence across the NHS workforce. The **second report** from this research will outline suggested pathways for related education and training.

Interviewees for this research identified broader efforts that primarily aim to improve patient safety and service delivery, but could also contribute to developing confidence in AI within the healthcare workforce.

Figure C shows these efforts mapped across this report's conceptual framework.

Much of this work is already underway, being led by Health Education England, the NHS Transformation Directorate, Integrated Care Systems and trusts, regulators and moderators, legal professionals, academics, and industry innovators.

A forthcoming project will involve engagement with these organisations and relevant groups and sharing of updates on progress being made on these efforts.

Figure C. Efforts that can contribute towards confidence in AI



Governance

- » Development of professional guidelines on creating, implementing, and using AI for all clinical staff groups
- » Further development of regulatory frameworks for AI performance, quality, and risk management
- » Finalisation of formal requirements for evidence and validation of AI technologies
- » Development of AI specific pathways for prospective clinical studies of new technologies
- » Further development of guidance on liability for AI (including autonomous AI)
- » Establishment of flexible and dynamic processes for developing clinical guidelines on AI-assisted clinical tasks and technologies
- » Development of clear oversight and governance pathways for AI, including AI not classified as a medical device
- » Development of standards for developing AI for health and care settings (including co-creation with users, model transparency and mitigation of model bias)



Implementation

- » Further development of advice, guidelines, and prototypes for information technology (IT) and governance (IG) supporting adoption of AI technologies
- » Development of strategies and assignment of resources to encourage organisational cultures that support innovation, co-creation, and robust appraisal of AI technologies
- » Encouragement of collaboration and sharing of knowledge across NHS sites that are adopting AI technologies
- » Development and resourcing of multi-disciplinary teams across clinical, technical, and administrative roles to enable implementation, local validation, audit and maintenance of AI technologies
- » Establishment of pathways for ongoing monitoring, performance feedback and safety event reporting involving AI technologies

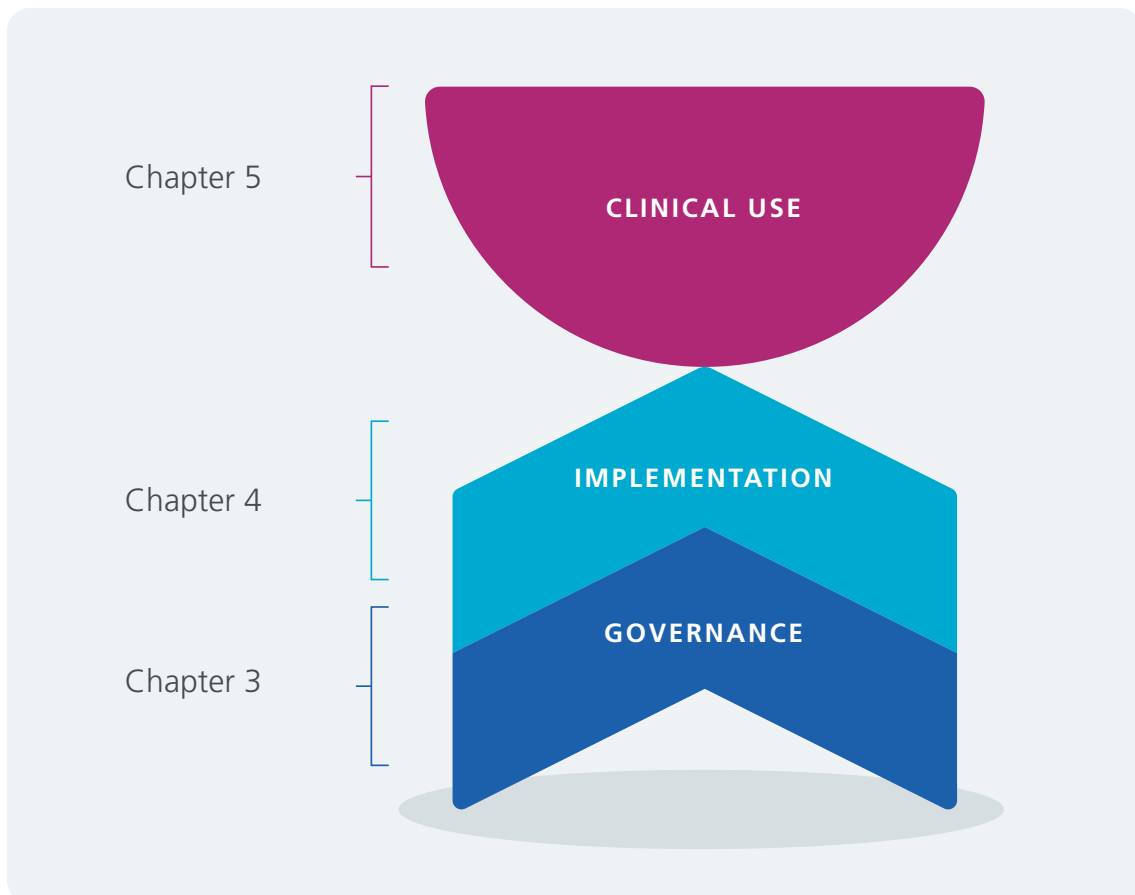


Clinical Use

- » Development of internal systems to record AI-assisted CRDM, including how AI has influenced or changed the decision
- » Further research on explainable AI and its safe use in clinical reasoning and decision making (CRDM)
- » Further research to understand and optimise the presentation of AI-derived information for CRDM
- » Further research to understand how certain AI model features influence confidence
- » Development of confidence in AI technologies across patients and communities via engagement and education activities
- » Support for clinicians to determine appropriate confidence in AI-derived information and balance it with conventional clinical information for CRDM

Report overview

- Chapter 1** **Chapter 1** provides the methodology and context for this research.
- Chapter 2** All readers are encouraged to read **Chapter 2**, which provides the key concepts and the conceptual framework for understanding confidence in AI among healthcare workers.
- Chapter 3** For readers involved or interested in AI-related governance (regulation, evaluation, guidelines, liability).
- Chapter 4** For readers involved or interested in the implementation of AI in local healthcare settings.
- Chapter 5** For readers using AI directly in clinical practice, or managing and educating such users.





Chapter 1: Introduction

1.1 Research purpose

This research – a collaboration between Health Education England and the NHS AI Lab at the NHS Transformation Directorate – has its origins in the Topol Review (2019),⁵ which explored how to prepare the UK’s healthcare workforce to master digital technologies for patient benefit. The Topol Review recommended that the NHS should develop a workforce able and willing to transform it into a world leader in the effective use of healthcare Artificial Intelligence (AI) and robotics.

Health Education England (HEE) has since established the Digital, AI and Robotics Technologies in Education (DART-Ed) programme to understand the impact of advances in AI on education and training needs. This research will build on the AI Roadmap,³ published by HEE and Unity Insights (formerly the analytics and evaluation function of Kent Surrey and Sussex Academic Health Science Network) in January 2022. The Roadmap provides an understanding of the use of AI and data-driven technologies that currently exist in the healthcare system, the uptake of these new technologies, and the impact on the workforce. In addition, HEE is working with the University of Manchester to develop a skills and capabilities framework to support curriculum review and guide healthcare practitioners towards future required learning for digital healthcare.

The primary aim of this research is to explore the factors influencing healthcare workers' confidence in artificial intelligence (AI) technologies and how these can be addressed through education and training.

Supporting healthcare workers to feel confident in identifying when and how to use AI is a main objective of the NHS AI Lab, and a key component of its vision for the safe, effective, and ethical adoption of AI technologies across health and care. This research will support the Lab's commitment to empower healthcare workers to make the most of AI, including making the best of their expertise, informing their decisions, and saving them time to focus on patient care.

1.2 Methodology

The research involved a review of related academic literature, and semi-structured interviews exploring experiences of developing and using AI technologies in healthcare settings.

Interviewees included workers in primary and hospital care settings with varying levels of experience in AI technologies; industry innovators; representatives of related regulatory and arm's length bodies; and academics who work at the intersection of AI, healthcare, education and clinical confidence.

The research did not include workers and carers in social and community care settings, although some of the findings may be relevant to aspects of their work.

Appendix A provides a list of the individuals and organisations interviewed for this research.

1.3 Research reports

This first report provides an analysis of the literature review and the interviews conducted for this research, synthesised into a conceptual framework for understanding confidence in AI among healthcare workers.

A **second report** will determine educational and training needs based on the findings and conceptual framework presented in this first report. It will outline pathways to inform how educational providers, industry innovators, and healthcare providers can develop related education and training offerings for current and future healthcare workers.

Audiences for this report



The report aims to enhance understanding of what influences confidence in AI among healthcare workers. It can be used by policymakers, regulators, arm's length bodies, legal professionals and industry innovators to guide activities that can contribute towards developing confidence in AI. These activities, many of which are already underway, are listed in the **Conclusion** section.



The report can also be used as guidance by healthcare workers who want to understand concepts related to confidence for AI, to inform how they adopt, implement, and use AI technologies in their settings. These can include individuals responsible for strategic decisions and for the procurement of AI technologies, as well as regular users of these technologies.

1.4 Context

AI technologies are the latest innovation in a series of digital technologies that have been transforming the delivery of healthcare. The broader efforts to enable the adoption of change and innovation in health and care settings (such as the digital transformation pathways outlined in the NHS Long Term Plan)⁶ are important foundations for the adoption of AI technologies.

The UK has published an ambitious strategy to remain a science and AI superpower in the next ten years.⁷ The NHS is supporting this objective by developing a strategy for AI in health and care.⁸ Building on the work of the NHS AI Lab, the strategy's vision is to enable the safe scaling of proven and fair AI technologies that deliver better outcomes for the UK population.

Existing levels of familiarity and experience in AI technologies

International surveys have shown that most healthcare workers lack direct experience with AI technologies.⁹ In the UK, a 2020 survey of over 1,000 NHS staff by the Health Foundation² found that three-quarters of respondents have heard, seen or read 'not very much' or 'nothing at all' about automation and AI.

The survey's respondents were split between being positive or negative overall about the use of AI in healthcare, with some notable insights:

- » those more familiar with AI technologies tended to be more positive towards these technologies.

- » medical and dental staff respondents felt more positive about AI than nurses and midwives who in turn felt more positive than health care assistants.
- » assistive applications of AI, like image analysis and screening, were perceived as a bigger opportunity than autonomous forms of AI like robotic care assistants.

Respondents perceived the main benefits of automation and AI would be to improve efficiency and free up time to care for patients, a key objective identified in the Topol Review. The biggest identified risks of using automation and AI involved healthcare becoming more impersonal and healthcare workers failing to question any suggestions or decisions proposed by these systems. The biggest identified challenges were patients potentially not accepting these technologies and being suspicious of them, and staff shortages or inadequate equipment for implementation and use.

While most of the Health Foundation survey's respondents anticipated that AI would improve their quality of work, a notable subset viewed AI as a threat to their jobs and professional status. This fear of AI replacing human jobs has been noted in several other research studies (particularly in relation to administrative positions and in radiology and pathology) along with other concerns like data governance, cyber security, patient safety, and fairness.^{5,10,11}

The healthcare workers interviewed for this research shared similar views about the levels of experience and attitudes towards AI technologies within their settings, noting that most related activities involve a 'small pool of experts' employed by the NHS who are often self-taught. They noted that exposure to AI technologies varies across professions and roles, with radiology, dermatology, ophthalmology, and pathology leading in this space (driven by the need for, and availability of, AI-enabled machine vision technologies).

However, interest in and understanding of how AI technologies can assist ongoing clinical and administrative challenges and improve patient outcomes is growing. Interviewees discussed current work in their settings on how to coordinate or develop data sets to train algorithms, investigations around static and dynamic AI models, and concerns about bias in AI systems. They also highlighted perceived challenges in ensuring the ongoing robust performance of AI systems and in understanding their impact on clinical safety and real-world patient outcomes.

The growing deployment of AI in health and care settings is highlighted also in HEE's AI Roadmap report,³ which surveyed over two hundred AI technologies nearing or ready for market. The survey found that over 20 per cent of these technologies are estimated to be ready for large scale deployment within 2022, and an additional 40 per cent in the next three years. The most affected workforce groups identified in the survey included radiologists, general practitioners, workers in non-clinical administration, diagnostic radiographers, and cardiologists.

1.5 Terminology

This report uses the terms '**AI**' and '**AI technologies**' to describe the use of digital technologies to create systems capable of performing tasks commonly thought to require intelligence. These can include algorithms using statistical techniques that find patterns in large amounts of data, or to perform repetitive cognitive tasks with data without the need for constant human oversight.

This definition of AI is intentionally broad and could encompass algorithms not commonly considered as AI. While parts of this report refer only to more complex machine-learning algorithms, many of the factors that influence confidence described in the report could apply to any data-driven technology or algorithm used in healthcare or clinical practice.

AI technologies have the potential to support existing clinical capabilities in diagnosis and screening, drug discovery, digital epidemiology, and personalised medicine,¹ as well as optimising organisational resources, system efficiencies and clinical workflows.

Clinicians, as referred to in this report, include healthcare workers making a patient-specific decision that affects patient care, and may include Nurses, Paramedics, Allied Health Professionals, Doctors, and other specialist healthcare staff groups.

Industry innovators refer to private sector developers and providers of AI technologies.



Chapter 2: Key Concepts and Framework

This chapter provides an overview of the key concepts and the conceptual framework for this report.

2.1 Key concept: Confidence in AI

Interviewees for this research noted that many of the UK healthcare settings that are currently adopting AI technologies are at a critical juncture. They are moving through the early stages of an AI technology's development (from the proof-of-principle to proof-of-efficacy stages) and introducing AI technologies in clinical trials.

Currently, the main challenges of these rollouts involve considerations around information technology (IT) systems, interoperability, and information governance. Interviewees noted that, as these initial challenges are resolved, issues relating to workflow integration, performance monitoring, demonstrating evidence of safety and effectiveness, and securing the trust and confidence of the workforce in AI technologies will become prominent.

The latter challenge is the focus of this research, starting with clarifying what trust, trustworthiness and confidence mean at the intersection of AI and healthcare.

Moving from trust to appropriate confidence

The literature review and analysis of the interviews conducted for this research suggest that trust and confidence are often used interchangeably, with increased trust in AI often stated as a desirable objective in health and care settings.

Therefore, when considering how AI is used in health and care settings, it is important to differentiate between trust (which is placed in a product or system), trustworthiness (which is earned), and confidence (which is held individually or collectively). These distinctions have informed this report's conceptual framework, which uses the term confidence rather than trust.

The term **trust** is a belief in a product or system and is typically a binary concept such that the subject is either trusted or it is not.

Trustworthiness encompasses the quality of a product or system being deserving of trust or confidence.

Confidence, like trust, conveys a belief in a product or system. However, unlike trust, it is not generally considered a binary concept. Instead, confidence can be understood as continuously variable and depending on various factors. Confidence can account for the nuances of using AI clinically, where higher confidence in AI-derived information is not always a desirable objective. It allows for a more dynamic exploration of related influences and behaviours, including where lower confidence may be justified.


Interviews for this research suggest that confidence in any AI technology or system used in health and care can be increased by establishing its trustworthiness. Increasing confidence for this purpose is desirable and can be accomplished through a multifaceted approach including regulatory oversight, real-world evidence generation and robust implementation.⁴

In the context of clinical decision making, once trustworthiness in AI technologies has been established, high confidence in AI-derived information (the output provided by an AI system to a clinician) may not always be desirable. Instead, different levels of confidence may be held in individual outputs from a given AI technology, depending on the context and circumstances. During clinical decision making, confidence in AI-derived information will depend on numerous factors including the clinical scenario and other available sources of information. The challenge, therefore, is to enable users to make context-dependent value judgements and continuously ascertain the **appropriate level of confidence in AI-derived information**, balancing AI-derived information against conventional clinical information.

2.2 Key framework: Understanding confidence in AI

This report presents a framework for understanding what influences confidence in AI within health and care settings, which was developed following analysis of the academic literature and the interviews conducted for this research.

Establishing confidence in AI can be conceptualised as:

-  » **Increasing confidence in AI by establishing its trustworthiness (applies to all AI used in healthcare)**
 - › Trustworthiness can be established through the **governance** of AI technologies, which conveys adherence to standards and best practice, and suggests readiness for implementation.
 - › Trustworthiness can also be established through the robust evaluation and **implementation** of AI technologies in health and care settings.
 - › Increasing confidence is desirable in this context.


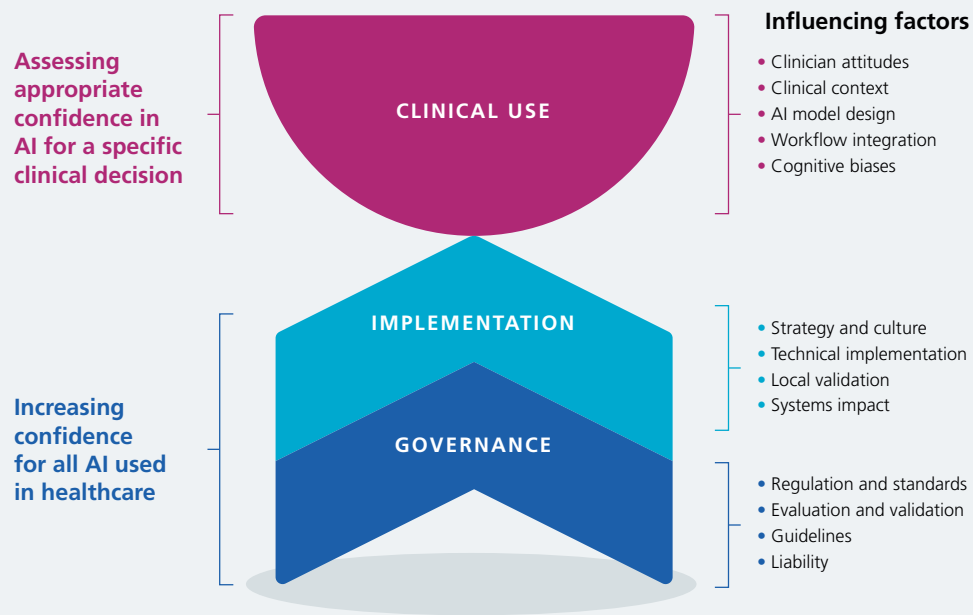
-  » **Assessing appropriate confidence in AI (applies only to AI used for clinical decision making)**
 - › During clinical decision making, clinicians should determine appropriate confidence in AI-derived information and balance this with other sources of clinical information.
 - › Appropriate confidence in AI-derived information will vary depending on the technology and the clinical context.
 - › High confidence is not always desirable in this context. For example, it may be entirely reasonable to consider a specific AI technology as trustworthy, but for the appropriate confidence in a particular prediction from that technology to be low because it contradicts strong clinical evidence or because the AI is being used in an unusual clinical situation. The challenge is to enable users to make context-dependent value judgements and continuously ascertain the **appropriate level of confidence in AI-derived information**.

Figure 1 illustrates the conceptual framework and lists corresponding factors that influence confidence in AI. These comprise factors that relate to governance and implementation, which can establish a system's trustworthiness and increase confidence. Clinical use factors affect the assessment of confidence during clinical decision making on a case-by-case basis.

Chapters 3, 4 and 5 explore the factors in detail, including how these are supported by current initiatives and guidance.

Figure 1: Framework for understanding confidence in AI among the healthcare workforce



The primary focus of this report is understanding and assessing appropriate levels of confidence in AI-derived information during clinical use. However, as appropriate confidence in AI used in clinical decision making is premised on establishing the trustworthiness of these technologies, the key elements of governance and implementation that underpin confidence are addressed first before clinical use is discussed in more detail.

Governance

Increasing confidence through the governance of AI technologies

How AI technologies are governed can influence confidence in these technologies.

Formal means of governance and oversight can increase confidence in AI among workers who plan, implement and use AI for any task in health and care settings. These can include robust regulatory frameworks and standards for AI, clear evaluation and validation approaches, clinical and technical guidelines, and clarity on liability across different AI technologies.

Implementation

Increasing confidence through the robust implementation of AI technologies

The safe, effective, and ethical implementation of AI in health and care settings are key contributors to confidence in AI technologies among the workforce.

The leadership, management and governance bodies within health and care settings can support such implementation by establishing AI as a strategic asset, such as through developing business cases, and maintaining organisational cultures conducive to innovation, collaboration, and public engagement (**Section 2.4** discusses the importance of developing confidence in AI among patients).

Addressing any challenges with information technology infrastructures, interoperability, and data governance requirements is also crucial. Establishing and agreeing on related information technology and governance arrangements are instrumental to healthcare workers' confidence in using AI technologies.

Procurement or commissioning entities within health and care settings will need to decide whether to validate the performance of AI technologies to ensure its performance translates to local data, patient populations and clinical scenarios. In such cases, evaluation using local data and workflows will be a necessary step to the robust implementation of AI technologies.

Healthcare workers will be more confident in AI technologies that are safely and efficiently integrated into existing workflow systems, including through established pathways for reporting safety events.

Clinical use

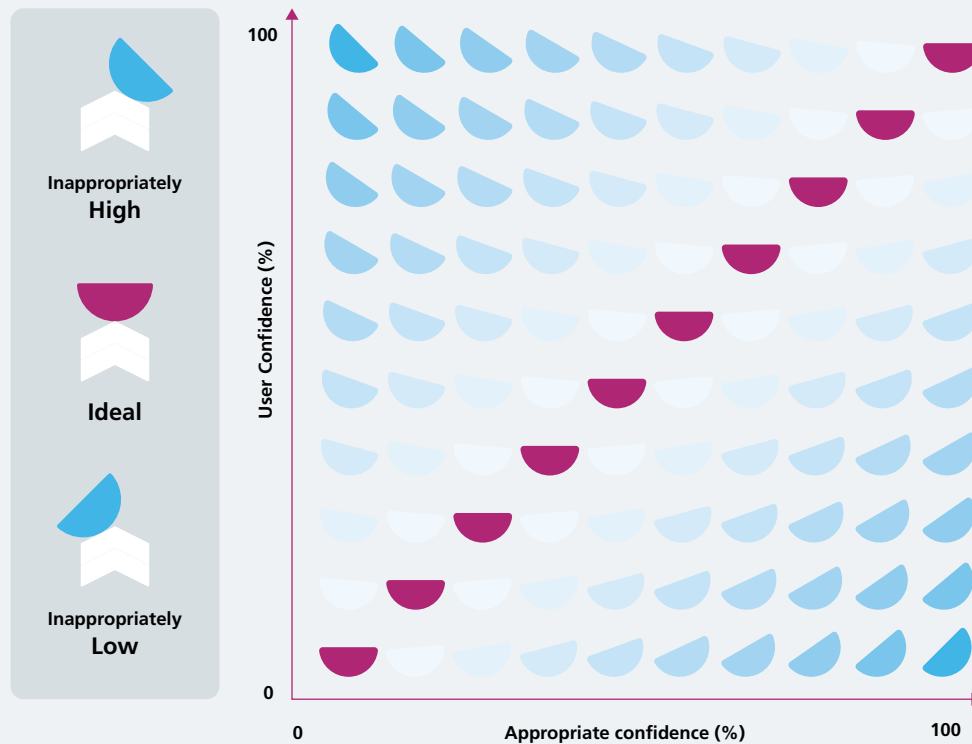
Assessing appropriate confidence in AI-derived information during clinical decision making

Figure 2 shows that a clinician's actual confidence (shown as 'User confidence') in AI-derived information for decision making may be inappropriately high or low if it does not match (along the 'ideal' line) the appropriate level for a clinical case. As discussed in this section, this appropriate level is likely to vary from case to case, depending on clinical factors and the AI-derived information itself.

To avoid inappropriately high or low levels of confidence, clinicians need to determine the appropriate level of confidence in the specific AI-derived information available at the point of making each clinical decision.

A complex set of considerations can dictate how to determine an appropriate level of confidence in AI-derived information, depending both on the technology and the clinical scenario, and with certain AI technologies and use-cases presenting lower clinical or organisational risks.

Figure 2: Ideal and inappropriate levels of confidence



Synthesising and evaluating information from many disparate sources are key skills in clinical decision making, whether for diagnosis, prognostication, or treatment. If incorporated and considered with appropriate confidence, information from AI technologies has the potential to make clinical decision making safer, more effective, and more efficient.

To achieve this, clinicians will need to understand when AI-derived information should and should not be relied upon, and how to modify their decision making process to accommodate and best utilise this information. This might include considering factors like:

- » other sources of clinical information and how to balance these with AI-derived information in decision making
- » the clinical case for which the AI is being used
- » the intended use of the AI technology.

Several factors can influence how clinicians view AI-derived information (their user confidence), potentially leading to inappropriately high or low levels of confidence. These include:

- » Clinicians' personal experiences and **attitudes**. General digital literacy, familiarity with technologies and computer systems in the workplace, and past experiences with AI or other innovations can influence assessments of confidence in AI-derived information.
- » **Clinical context** including the level of clinical risk and the degree of human oversight in the AI decision making workflow.
- » **Characteristics** of AI model design. Various design characteristics can influence confidence in AI technologies. For example, the way AI predictions are presented (such as diagnoses, risk scores, or stratification recommendations) can affect how clinicians process information and potentially influence their ability to establish appropriate confidence in AI-derived information.
- » **Cognitive biases**, including automation bias, aversion bias, alert fatigue, confirmation bias and rejection bias can affect AI-assisted decision making. The propensity towards these biases may be affected by choices made about the point of integration of AI information into the decision making workflow, or the way such information is presented. Interviewees for this research highlighted that enabling clinicians to recognise their inherent biases and understand how these affect their use of AI-derived information should be a key focus of related training and education. Failure to do so may lead to unnecessary clinical risk or the diminished patient benefit from AI technologies in healthcare.

Clinicians will need to understand when AI-derived information should and should not be relied upon, and how to modify their decision making process to accommodate and best utilise this information. Awareness of how their own attitudes and cognitive biases, the clinical context, and AI technical features can influence how they use AI-derived information will be crucial to ensure appropriate levels of confidence.

2.3 The importance of developing confidence in AI among healthcare workers

Interviewees for this research stressed the importance of the healthcare workforce being confident in adopting AI technologies.

Low confidence may limit the use of AI technologies and result in wasted resources, workflow inefficiencies, substandard patient care and potential disparities in who gets to benefit from AI technologies.

During clinical decision making, inappropriate levels of confidence in AI-derived information could lead to clinical errors or harm in scenarios where the AI underperforms, without being properly assessed or checked. This includes a phenomenon known as automation bias where the user inappropriately favours suggestions made by automated decision making systems.

As discussed in **Box 1**, maintaining appropriate confidence in AI-derived information is fundamental to the safe, effective, and ethical adoption of AI across health and care.

Box 1. Appropriate confidence in AI-derived information and ethical AI

Ethical AI encompasses practices that aim to address the individual and societal harms AI might cause.¹²

In health and care settings, the optimal care of patients and avoidance of harm are paramount, as are endeavours to minimise disparities in patient outcomes between demographic groups, geographic locations and healthcare organisations.

A major limitation in how AI technologies are developed and deployed currently is the potential for negative impact towards certain patient groups, including through biases built into AI models.^{13,14}

Prevention and mitigation of these biases are critical aspects of ethical AI. Minimising bias and maximising AI performance in real-world settings can be achieved to some extent through using representative data sets and robust evaluation and implementation.

Understanding and critically appraising AI-derived information – essentially, maintaining an appropriate level of confidence – can further assist in identifying potential failure cases of AI technologies, including in relation to bias, which can, in turn, contribute to an ethical AI approach.

Appropriate levels of confidence in AI-derived information can also ensure trust is sustained in patient-clinician relationships. Clinicians will need to be able to explain their reasoning around the use of AI to their patients to maintain informed decision making and patient empowerment.

These factors suggest that critically appraising AI technologies is key to the ethical adoption of AI in health and care settings. Education and training will be essential to improve related knowledge and skills to avoid healthcare workers having inappropriately low or high confidence in AI. The **second report** outlines suggested pathways for such education and training.

2.4 The importance of developing confidence in AI among patients and the public

While this report focuses on understanding confidence in AI technologies among the healthcare workforce, it is important to recognise that improving confidence in AI amongst patients and the public (including through public engagement and participation initiatives) will play an equally crucial part in the successful adoption of AI technologies in health and care settings.

Interviewees for this research cautioned that without public acceptance and confidence in AI, it will be highly challenging from an ethical and patient-led care perspective to use these technologies in health and care settings. Developing confidence in AI technologies across patient groups and communities through public participation and education activities will be a necessary component of a holistic approach to deploying AI technologies effectively and safely.

A survey of people across England, conducted as part of developing the National Strategy for AI in Health and Social Care,⁸ found that almost half of the thousand respondents had heard nothing at all or very little about AI. The results also showed that greater awareness or understanding of AI leads to greater belief in the benefits it can bring to health and social care.

Interviewees for this research noted that conducting early and ongoing engagement of patients and the public, to inform how AI is developed and implemented, can enhance confidence in AI both amongst the public and in the healthcare workforce (see [section 4.1](#)).

Although most current activities focus on the provision and access of patient data, further activities could aim to strengthen the involvement of patients in the design, governance and implementation of AI, including in safety reporting and post-market surveillance (monitoring the performance and safety of an AI technology when released on the market).

For example, the NHS AI Lab will be trialling the engagement of patients to identify possible risks and biases during the early stages of AI product development. This will support industry innovators to assess these risks and make any necessary adjustments to their products.¹⁵



Chapter 3: Governance



This chapter details four main factors, as identified by interviewees for this research, that relate to the governance of AI technologies and influence confidence in these technologies: regulation and standards; evaluation and validation; guidelines; and liability.

These factors underpin aspects of the trustworthiness of AI technologies, and entail robust, reliable and established direction and oversight from central healthcare leadership and the main regulators of healthcare in the UK (as shown in **Figure 3**). As described in this chapter, these efforts are currently in development and at different levels of progress.

Confidence in how AI is governed can enable clinicians to assess appropriate levels of confidence in AI-derived information during clinical decision making, as detailed in **Chapter 5**.

3.1 Regulation and standards

Confidence that AI technologies are included in formal governance and oversight

Interviewees for this research highlighted that a robust, efficient and transparent regulatory system can support confidence in the safe and effective adoption of AI technologies in health and care settings. This includes the regulation of AI products, the regulation of healthcare settings and the regulation of healthcare professionals.

Navigating the regulatory landscape for AI technology can be complex and confusing for both industry innovators and adopters of AI. To simplify this, key UK regulatory and arm's length bodies (the National Institute for Health and Care Excellence, the Medicines and Healthcare products Regulatory Agency, the Health Research Authority, and the Care Quality Commission) are developing MAAS (Multi-Agency Advice Service), a cross-regulatory advisory service for developers and adopters of AI.¹⁶ MAAS will create educational material about the regulation of AI technology and provide access to the information developers and procurers of AI need to ensure products are meeting regulatory requirements.

3.1.1 Regulation of AI products

Regulation can support confidence in AI products used in health and care, giving assurance that it has been developed responsibly, works as advertised, and that patient data is used in a safe, secure and responsible way.

Regulatory requirements for AI products vary depending on whether an AI product is classed by the Medicines and Healthcare products Regulatory Agency (MHRA) as a medical device.

Medical devices must be registered with the MHRA and are subject to Medical Device Regulations, the UK MDR 2002. This regulation is supported by standards (for example, from the International Organisation for Standardisation) that can be used to demonstrate conformity with medical device regulation.

AI products used in health and care settings that are not classed as medical devices, such as products used to automate administrative processes, are not regulated by the MHRA. These products must however conform with other regulations like the General Data Protection Regulation (GDPR) and the NHS Digital Technologies Assessment criteria (DTAC) framework.¹⁷

Regulation of medical devices

Interviewees for this research suggested that there are gaps in the existing regulatory landscape for medical devices. They suggested that regulatory approval does not meet the expectations of AI users and that additional, AI-specific, regulation may be required.

All medical devices, including software as a medical device (SaMD), marketed in the UK must be registered with the MHRA and comply with the UK MDR 2002. UKCA certification is required for a device to be placed on the UK market (CE marking, the EU equivalent, will no longer be valid after 30 June 2023).

Devices are classified in accordance with UK MDR 2002 based on their clinical risk (Class I, IIa, IIb and III) with higher classifications associated with higher clinical risk and more stringent regulatory requirements.

Interviewees for this research suggested that there is very limited understanding of medical device and software regulation amongst the healthcare workforce. For example, interviewees perceived that most NHS professionals are unaware of what classifies a product as a medical device or the distinctions between classes of medical device, and what this means for product assessment and deployment.

In addition, interviewees perceived that healthcare workers often equate regulatory approval with proof that a product has met certain standards and can be trusted, for example, that it works in real-world clinical settings. However, current regulatory standards may not provide the assurances that healthcare workers assume they do. In particular, MDR compliance is focussed on quality systems for recording design decisions and testing processes, rather than clinical or technical evidence of performance. The performance of AI technologies requires a different evaluation from other medical devices. This suggests the importance of the workforce understanding the remit of regulatory approval, and clarifying what it does and does not guarantee for a given AI technology.

The regulation that currently applies to AI medical devices is the same regulatory framework that is used for any SaMD. Many of our interviewees felt that tailored AI regulation may be necessary to address AI-specific risks.

These developments are already in progress. In September 2021, the MHRA announced the Software and AI as a Medical Device Change Programme,¹⁸ which includes three packages specific to AI as a medical device:

- » Project AI RIG (AI Rigour) – to ensure AI is safe, effective and fit for purpose for all populations that it is intended to be used on.
- » Project Glass Box (AI Interpretability) – to outline the impact of interpretability on the safe and effective development and use of AI medical devices.
- » Project Ship of Theseus (AI Adaptivity) – to create guidance that allows for adaptive AI that does not fit within existing change management processes.

Standards

The UKCA marking and approval of SaMD is dependent on conformance with the UK MDR 2002, which requires manufacturers to maintain quality management systems. One way of demonstrating conformance is to meet a 'designated standard'.¹⁹

Box 2. ISO standards

The standard 'ISO 13485:2016 Medical devices - Quality management systems - Requirements for regulatory purposes' specifies requirements for a quality management system to demonstrate a manufacturer's ability to provide medical devices and related services that consistently meet customer and applicable regulatory requirements. The 'ISO 14971 Medical devices — Application of risk management to medical devices' standard provides further details on risk assessment, control, review and monitoring.

While it is not mandatory to follow the ISO 13485 standard, it is an effective option to demonstrate compliance with UK MDR 2002 in quality management and to follow internationally recognised best practice. For example, the UKCA marking is based on the requirements of the EU MDD (Directive 93/42/EEC), which states in clause 12.1a *'For devices which incorporate software or which are medical software in themselves, the software must be validated according to the state of the art taking into account the principles of the development lifecycle, risk management, validation and verification.'* The ISO 13485 standard can guide these aspects of medical device development, detailing also the requirements for validation.

ISO 13485 mentions software explicitly, following the categorisation of SaMD guidance published by the International Medical Device Regulators Forum (IMDRF).²¹ ISO 13485 does not differentiate SaMD from conventional hardware medical devices in terms of requirements for quality management but recognises there may be differences in the way in which the requirements are met.

Although SaMD is recognised in ISO13485, AI as a medical device (AIaMD) is currently not mentioned in standards surrounding medical devices, leading to the need to interpret the requirements for this context. Standards and guidance accompanying the MHRA Software and AI as a Medical Device Change Programme will aim to clarify the specific requirements for AIaMD.¹⁸

The usability elements of the ISO 9241 standard are also applicable to AI in healthcare, describing user interface and experience principles for human-system interaction. Part 810 of the standard discusses the usability of 'Robotic, intelligent and autonomous systems', highlighting some of the system complexity and human-system interaction challenges relevant to AIaMD. Compliance with this standard is not currently a requirement of UKCA or CE marking but could be considered best practice towards building confidence in AI technologies.

As detailed in **Box 2**, International Organisation for Standardisation (ISO) standards can provide a framework for manufacturers to demonstrate the suitability of their product design and quality management systems.

In order to implement software within the NHS, all digital health products must also comply with NHS digital, data and technology standards.²⁰ These include DCB 0129 and DCB 0160, which set out the clinical risk management framework for health organisations and suppliers of digital technology used in the health and care environment.

3.1.2 Regulation of healthcare settings

While medical devices are regulated by the MHRA, healthcare settings are regulated by the Care Quality Commission (CQC).

The CQC monitor and inspect services and assess whether they are safe, effective, caring, responsive and well led.²² They publish standards of care, setting out what good and outstanding care looks like and make sure services meet fundamental standards below which care must never fall.²³

In the context of AI-enabled services, the CQC's role includes ensuring healthcare settings meet fundamental standards of quality and safety during an inspection, regardless of the medical device status of the technology used.

The CQC has published principles for the inspection process for particular types of technology such as surveillance CCTV.²⁴ These principles assess whether surveillance technology is safeguarded, secured, lawful, transparent, operated by trained staff and used in a manner that maintains patient involvement, privacy and dignity.²⁵ Principles for the safe and effective use of AI technologies may also be appropriate.

3.1.3 Regulators of healthcare workers

Regulators of healthcare workers, like the General Medical Council (GMC) and the Nursing and Midwifery Council (NMC), are responsible for setting standards of competence and conduct, and assessing the quality of education and training courses to ensure healthcare workers have the skills and knowledge to practise safely and competently. These standards may need to be revisited in the context of AI technologies.

Interviewees for this research noted that clinicians look to regulators for guidance on how they should use AI technologies and for reassurance that using AI in clinical practice will not threaten their professional registration. Therefore, the position regulators take about AI technologies will significantly influence clinician confidence in these technologies.

The General Medical Council (GMC) and Health and Care Professionals Council (HCPC) codes of conduct require that clinicians must be prepared to explain and justify their decisions.^{26,27} This may be challenging in situations in which 'black box

AI' is used in clinical decision making where the clinician cannot explain how an algorithm has reached a given conclusion.

Regulatory standards apply both to clinicians using AI in clinical decision making and those involved in the design, testing and validation of AI products. Interviewees noted that the latter are undertaking roles that were not traditionally within the remit of regulators of healthcare workers, and may require particular consideration and specialised guidance.

Non-clinical developers of AI products used in healthcare are not regulated in the same way as clinical professionals. Feedback from the interviews conducted for this research suggests that formal registration and accreditation for these roles by a regulatory body may be beneficial. This might include a systemised set of training protocols including technical, ethical and safety standards. A formal accreditation could act to promote the development of safe and effective AI and improve public trust in these technologies.

Key confidence insights

- » A robust regulatory system is key to ensuring that technologies are safe and effective, which contributes to the trustworthiness of AI systems.
- » Healthcare workers may assume regulatory approval proves that an AI product works in a real-world clinical setting. However, current regulatory standards do not provide this level of assurance regarding performance.
- » The healthcare workforce will need to understand the remit of regulatory approval of medical devices.
- » Principles for the safe and effective use of AI technologies from regulators of healthcare settings may be appropriate.
- » Regulators of healthcare workers can consider how to advise clinicians who develop, validate and use AI technologies.
- » Formal registration and accreditation of non-clinical developers of healthcare AI products may be beneficial to promote the development of safe and effective AI.



3.2 Evaluation and validation

Confidence that AI technologies work in real-world clinical settings

This section discusses the required evidence and the processes used to determine how well an AI technology performs according to its intended use (referred to as efficacy). Interviewees for this research highlighted that these contribute to establishing the trustworthiness of, and hence increasing confidence in AI technologies.

Formal requirements for evidence in AI technologies are still being developed.

Evidence of an AI's efficacy can be captured in several stages. AI technologies can progress through these stages during their development and deployment:

- 1. Internal validation:** The AI model is tested by its developer using a separate validation data set, often split from the same source as the training data set. It generally uses retrospective data sets (data that has been collected in the past).
- 2. External validation:** The AI model is tested with data from a different source to the training data set. It tests the generalisability of the model's performance to clinically relevant scenarios, ensuring the performance that has been validated internally is maintained. External validation may be performed by the AI developer, or independently by a third party.
- 3. Local validation:** In some cases, limited further external validation may be needed as part of deploying AI at a local setting, to ensure its performance translates to the local data, patient populations and clinical scenarios. Procurement or commissioning entities may decide that the available evidence for an AI technology does not provide sufficient confidence that its performance will be acceptable in the local situation (Local validation is discussed in [section 4.3](#)).
- 4. Prospective clinical studies:** The AI model is tested in a real-world clinical setting using data collected in real time. This evaluation determines whether the technology has benefits in terms of efficiency or patient outcomes. It involves testing the AI's technical performance as well as its integration with clinical workflows.
- 5. Ongoing monitoring:** Healthcare settings should monitor the performance of AI algorithms in use to ensure there is no degradation due to population drift or technical factors elsewhere in the data pipeline.

Research has shown that many AI models that perform well at the internal validation stage perform significantly worse at the external validation stage.²⁸ This may be due to the model being insufficiently 'generalisable'; in other words, the model does not replicate its success when given data different to its original source.

Further, products that perform well on both external and internal validation can perform poorly in prospective clinical studies.²⁹ There can be many reasons for this poor performance, including human factors, technology infrastructure, and systems considerations.

Current approaches

Interviewees for this research noted that uncertainty in the appropriate evidence required for AI technologies can have a negative impact on the confidence of those commissioning and using AI technologies.

Current MHRA guidelines for UKCA regulatory approval stipulate the need for internal validation testing. Although the requirement for external validation (using data from a different source to the training data) may be inferred from the UKCA clinical evaluation requirements, there is currently no requirement for this validation to be conducted by an independent third party.³⁰ Prospective clinical studies are not currently a requirement for regulatory approval for SaMD.

Interviewees for this research, supported by literature, observed that although most AI technologies being considered by NHS settings have gone through internal validation, few have been externally validated by third parties or undergone prospective clinical studies.^{31,32} Interviewees observed that very few industry innovators are committed to long-term prospective clinical studies (which demand time and monetary resources). Instead, many present regulatory certifications (such as UKCA or CE certification) as evidence of their technology's performance.

Procurers of AI technologies and academics interviewed for this research suggested that the use of regulatory certification as evidence of product efficacy is inadequate for algorithms used in clinical decision making. They recommended that external validation and prospective clinical studies be required for UKCA approval, with evidence made publicly available.

Additional evidence is already required for AI used in national screening programmes. The UK National Screening Committee stipulates that AI screening technology must undergo external validation with further clinical evaluation being required in certain circumstances.³³

Interviewees, supported by research, highlighted that prospective clinical studies are a particularly important driver for establishing confidence in AI technology that directly impacts clinical decisions or patient outcomes.^{9,34} These prospective studies can identify the system benefits, such as saving money, freeing up staff time for other tasks, or enabling a shift from inpatient to outpatient care.

Clinician interviewees for this research perceived that AI technologies used in patient care should be held to the same evidence standards as other medical interventions, such as pharmaceuticals. This includes systematic reviews, randomised controlled trials (RCTs) and peer review research. Revision and redesign of existing information technology (IT) and data governance infrastructures may be needed to support this level of evidence generation.

On the flip side, industry innovators interviewed for this research expressed concerns that traditional prospective clinical study methods such as RCTs require significant financial investment and long-time scales. They suggested that alternative evidence generation methods could be considered to ensure the healthcare system can benefit from new technologies sooner rather than later.

Several standards and tools have or are being developed for medical devices and clinical research to address the issues of varying evidence standards for AI products.

Box 3 provides a few examples.

Box 3. Evaluation and validation standards and tools

The National Institute for Health and Care Excellence (NICE) evidence standards framework

The NICE evidence standards framework³⁵ for digital health technologies is being updated to include AI-specific evidence guidelines dependent on the level of clinical risk associated with the technology. These aim to demonstrate that AI technologies are clinically effective and offer economic value, and could become the benchmark of evidence for AI technologies used in the NHS.

META tool

The META (MedTech Early Technical Assessment) Tool³⁶ has been developed by NICE in collaboration with Health Innovation Manchester. It is a tool that can be used by developers of AI technologies to identify the gaps in evidence generation required to demonstrate their value to the NHS. It can help products progress to a successful NICE Medical Technologies Evaluation Programme or Diagnostics Assessment Programme application.

HealthTech connect

HealthTech Connect is a secure online system managed by NICE, with funding from NHS England and NHS Improvement, which is designed to identify and support health technologies as they move from inception through to adoption. A range of partner organisations contribute including the AHSN Network, Office for Life Sciences, the MHRA, industry associations (ABHI, AXREM, BIVDA), the National Institute for Health Research (NIHR) and NHS Clinical Commissioners. The system helps companies to understand what information is needed by decision-makers in the UK health and care system (including levels of evidence) and clarifies possible routes to market access.

Box 3. (cont.)***AHSN innovation exchange***

The AHSN (Academic Health Science Network) innovation exchange connects researchers, the life sciences industry and healthcare. One of its four structured elements relates to the validation of AI products in clinical settings. Through this process, AHSNs are developing a consistent approach to validate AI technologies.

AI evaluation and clinical trial reporting guidelines and tools

Several guidelines and tools have been developed for conducting and reporting external evaluation studies and prospective clinical trials for AI technologies. These are the result of an international collaborative effort to improve the transparency and completeness of reporting of evaluations and clinical trials of AI interventions. A few examples, some currently in development, include:

- **TRIPOD-AI**³⁷ (Transparent Reporting of a multivariable prediction model of Individual Prognosis Or Diagnosis-Artificial Intelligence)
- **STARD-AI**³⁸ (Standards for Reporting of Diagnostic Accuracy Study-AI)
- **PROBAST-AI**³⁷ (Prediction model Risk Of Bias ASsessment Tool-AI)
- **SPIRIT-AI**³⁹ (Standard Protocol Items: Recommendations for Interventional Trials-AI)
- **CONSORT-AI**⁴⁰ (Consolidated Standards of Reporting Trials-AI)
- **DECIDE-AI**⁴¹ (Developmental and Exploratory Clinical Investigation of DEcision-support systems driven by Artificial Intelligence)
- **QUADAS-AI**⁴² (Quality Assessment tool for artificial intelligence-centered Diagnostic test Accuracy Studies)
- **STANDING** together⁴³ (STANdards for Data INclusivity and Generalisability)

The limits of universal approaches to evaluating and validating AI

Interviewees for this research noted that while evidence standards set at the national or international levels can provide confidence in product efficacy, they will not negate the need for local validation of AI technologies.

The performance of AI systems can be strongly dependent on the details of the training data cohort and its similarity to the local cohort and situation.⁴⁴ An AI technology can behave differently in a new context either in terms of technical performance or clinical impact. Therefore, when an AI technology is being deployed in a particular location or for a particular use case, additional local validation may be expected.

This suggests the need for local skills and capabilities to assess whether local validation of an AI technology is desirable and appropriate, and to conduct this validation where necessary. Some NHS Trusts have already established specialised teams to coordinate local validation of AI technologies as discussed also in [section 4.3](#).

Key confidence insights

- » The evidence and the methods used to determine how well an AI technology performs according to its intended use contribute to confidence in AI technologies. These include internal validation, external validation, local validation, and prospective clinical studies.
- » Prospective clinical studies are a particularly important driver for securing confidence in AI technologies that directly impact clinical decisions or patient outcomes.
- » Healthcare workers expect AI technologies to be held to the same evidence standards as other medical interventions such as pharmaceuticals.
- » Clear guidance is needed for evidence provided for AI technologies currently deployed in the NHS. The NICE evidence standards are being updated to include AI-specific guidance.
- » The use of regulatory certification as evidence of product efficacy is inadequate for algorithms used in clinical decision making.
- » To support evidence generation, IT and governance infrastructures within the NHS will need to support prospective research/clinical trials that involve AI technologies.



3.3 Guidelines

Confidence that the right AI technologies are being procured and deployed

Interviewees for this research noted that procurement, ethical and clinical use guidelines can steer how AI is adopted and used within healthcare, and drive confidence in these technologies. They cautioned that effective guidelines would require a dynamic creation process to keep pace with AI development and adoption.

3.3.1 Procurement guidelines

Interviewees suggested that confidence in procuring suitable AI solutions within health and care settings is an important initial step to the safe and effective adoption of these technologies.

The NHS has developed initial guidance in its 'Buyer's Guide to AI in Health and Care'.⁴⁵ The guidance sets out important questions for public sector entities to consider when purchasing 'off-the-shelf' AI products (developed by industry innovators and packaged as ready for deployment). These include clarifications on the suitability of the solution, and regulatory, performance and ethical considerations.

The Digital Technology Assessment Criteria (DTAC) for health and social care can support further confidence in meeting clinical safety, data protection, technical security, interoperability and usability and accessibility standards.¹⁷

Disclosure

Industry innovators can be reluctant to share details of their AI products (including information on computational methods, and the robustness and completeness of their training data) due to commercial considerations and intellectual property rights. This can impact the confidence of those procuring, implementing and using AI technologies who may find it challenging to compare products, assess potential risks, determine the need for additional local validation, and communicate with patients about how the technology works.

Several significant developments in regulation and law relating to disclosure and transparency in AI models can guide future approaches to these challenges.

Box 4 summarises some of these initiatives.

Box 4. AI disclosure and transparency initiatives

GDPR transparency standards

The 2018 General Data Protection Regulation (GDPR) states that when automated decision making is used, the person to whom it relates should be able to access 'Meaningful information about the logic involved'. The guidance that accompanies the GDPR text gives further insights into the nature of this information, describing it as 'not necessarily a complex explanation of the algorithms used or disclosure of the full algorithm. The information provided should, however, be sufficiently comprehensive for the data subject to understand the reasons for the decision'. It will fall to legislators, data protection authorities, and courts to interpret when particular information will ultimately be classed as 'meaningful' and 'sufficiently comprehensive' without infringing on intellectual property rights.

CDDO algorithmic transparency standard

The Central Digital and Data Office (CDDO) recently announced an algorithmic transparency standard to help government departments provide clear information about the algorithmic tools they use, and why they're using them.⁴⁶

Model cards

Google have launched 'model cards' for AI algorithms - a structured way of sharing essential facts of machine learning models including their limitations.⁴⁷ 'Model Facts' labels specific to healthcare AI technology have also been created.⁴⁸

ICO and The Alan Turing Institute: Explaining AI in practice

The ICO and the Alan Turing Institute have released joint guidance on explaining decisions made with AI to give organisations practical advice to help explain the processes, services and decisions delivered or assisted by AI, to the individuals affected by them.⁴⁹

MHRA Project Glass Box

Project Glass Box (AI Interpretability) is one of the packages in the MHRA Software and AI as a Medical Device Change Programme.¹⁸ It aims to develop interpretability frameworks for AI algorithms to ensure they are sufficiently transparent to be robust and testable.

3.3.2 AI development guidelines

Guidelines for the development of AI can support industry innovators to create technologies in a safe and responsible way. Knowing that AI technologies have complied with these guidelines can also provide reassurance to those procuring and using these technologies.

There are currently several published initiatives to guide industry innovators and healthcare procurement, including:

- » the Department of Health and Social Care's 'Guide to good practice for digital and data-driven health technologies',⁵⁰ is designed to support industry innovators in understanding what the NHS is looking for when buying digital and data-driven technology for use in health and care
- » the NHS's What Good Looks Like framework⁵¹ is a guide for NHS leaders to digitise, connect and transform their services safely and securely. The framework involves seven success measures, including in relation to governance, resources and standards for safe care that can provide foundations for the development and deployment of AI technologies
- » the Office for Artificial Intelligence's 'Guide to using artificial intelligence in the public sector',⁵² outlines how to build and use AI in the public sector
- » 'Good Machine Learning Practice for Medical Device Development: Guiding Principles'⁵³ jointly published by the MHRA, The U.S. Food and Drug Administration (FDA) and Health Canada. These are ten guiding principles that can inform the development of medical devices that use artificial intelligence and machine learning.

3.3.3 Clinical guidelines

Clinicians often use guidelines to steer their diagnosis and management of patients.

Feedback from the interviews conducted for this research suggests that clinicians expect specific guidelines on the use of AI technologies to be developed and distributed by entities like the Royal Colleges and the National Institute for Health and Care Excellence (NICE). They consider these guidelines a key contributor to establishing their confidence in using AI technologies. The endorsement of an AI technology by a formal body like the Royal Colleges or NICE would be perceived as a key driver for adopting that technology.

However, the appropriate use of AI technologies will depend on specific features, like those outlined in **section 5.2**, as well as the clinical context. Interviewees cautioned that, while general guidelines may be sufficient for products with low clinical consequences, it is likely that specific guidelines will need to be developed for individual AI technologies that entail higher clinical consequences.

NICE has two programmes in which diagnostic technologies may be evaluated: the Medical Technologies Evaluation Programme (MTEP)⁵⁴ and the Diagnostics Assessment Programme (DAP).⁵⁵ Guidance has been published for several AI products.^{56,57}

Submission to MTEP or DAP is not a mandatory requirement for AI technologies and some industry innovators are uncertain about the suitability of these programmes for digital products. As documented in related research, innovators are deterred by the long timescales required to gather the necessary clinical trial evidence for the MTEP and DAP processes are at odds with the rapid iteration of digital technologies. They are concerned that their products may become outdated by the time they gain approval.⁵⁸

Further, interviewees for this research noted that smaller-size industry developers may not have the resources required to produce the level of clinical evidence required for their product's assessment.

NICE Medtech Innovation Briefings (MIBs) offer a faster way of obtaining NICE advice, taking around four months to produce. They do not provide the full guidance offered by MTEP and DAP but include a summary for the product, existing evidence, place in healthcare and expert opinion.^{59,60}

Interviewees for this research suggested that NICE guideline processes may be limited in scalability. The sheer volume and development lifecycle of AI technologies entering the market will potentially make it challenging for NICE to meet the demand for product-specific guidance. As multiple AI technologies for a given clinical task become available, it may be appropriate to move towards task-level guidance.

3.3.4 Ethical guidelines

The ethical dimensions of AI are currently being debated and defined, with some commonalities across the plethora of available frameworks. A worldwide study of related publications found universal inclusion of the principles of fairness and non-discrimination. Other prominent principles included privacy, accountability, and transparency.⁶¹

Although there is no universally adopted ethical guidance on AI, health and care settings and industry innovators can draw from available frameworks to inform their practices. These include:

- » the Central Digital and Data Office's 'Data Ethics Framework',⁶² which guides appropriate and responsible data use in government and the wider public sector
- » the World Health Organisation's 'Ethics and governance of AI for health',⁶³ which sets out six key ethical principles: protecting human autonomy; promoting human well-being and safety and the public interest; ensuring

transparency, explainability and intelligibility; fostering responsibility and accountability; ensuring inclusiveness and equity; and promoting AI that is responsive and sustainable.

Interviewees for this research highlighted the importance of developing awareness and recognition of ethical considerations like fairness, transparency, and accountability in health and care settings to complement regulatory and governance oversight.

Key confidence insights

- » Available guidelines can contribute to confidence in procuring, developing and using AI technologies. Knowing that AI technologies follow accepted guidelines can also contribute to confidence.
- » Clinicians may not feel confident using AI products in clinical decision making until they are included in established clinical guidelines.
- » Although general guidance may be helpful for AI products with low clinical risk, higher risk technologies are likely to require individual guidance. As more AI technologies enter the market, task-level guidance for AI technologies may be appropriate.
- » To support their confidence in AI, healthcare workers will need to develop awareness and recognition of ethical considerations like fairness, transparency, and accountability.

3.4 Liability

Confidence that attribution of liability is clear in relation to AI technologies

Interviewees for this research noted that clarity in the attribution of liability is crucial to increase confidence in, and enable the safe and ethical deployment of AI technologies.

They expressed concern about the current uncertainty in who will be legally accountable for AI technologies used in the clinical decision making process. They highlighted that establishing the liability of the various parties involved in designing, deploying and using AI will be important to promote confidence in these technologies.

Liability is a legal duty or obligation to take responsibility for one's acts or omissions. It is a longstanding legal principle that applies across sectors. However, it is unclear how liability will be applied to AI used in clinical decision making as there is a lack of established case law in this area.²⁷ The challenge of AI and liability is not unique to the healthcare industry and applies to various other sectors.

Responsibility for AI used in clinical decision making could fall to the clinician who uses the technology, the deploying organisation, the industry innovator that developed the technology or those who validated and approved the technology for clinical use. Various legal frameworks may be applicable including negligence, product liability and vicarious liability.

AI technologies that are used as a tool for decision making could feasibly be treated like other clinical decision making tools, with a potential focus on clinician accountability under medical negligence law. However, this may depend on the way in which an algorithm is used in the decision making process.

'Black-box' algorithms pose a particular challenge. If a clinician cannot fully understand and explain how a 'black-box' AI algorithm reaches its prediction, they cannot reasonably be considered accountable or responsible for the AI prediction itself.⁶⁴ However, they may still be held accountable for a decision made using the AI prediction.

In the case of autonomous AI, there is a potential that clinicians may be removed from the decision making process (for example, if AI were used to triage referrals or patient electronic consultations). If AI algorithms were fully responsible for a clinical decision in this manner it is unclear how existing legal frameworks would be applied.

The NHS AI Lab Futures Portfolio is looking at these issues in more detail, including through its 'Liability and Accountability programme', conducted with NHS Resolution and their expert legal panel, and a collaborative programme to assess the impact of meaningful human control in AI.

Clarity in liability will also influence developments in establishing related regulation and guidelines as discussed in **sections 3.1** and **3.3**.

Key confidence insights

- » Establishing the liability of the various parties involved in designing, deploying and using AI will be crucial to shaping confidence in AI.
- » Currently, there is uncertainty as to who will be held to account if AI products are used to make clinical decisions that lead to patient harm.
- » Clarity in liability will influence developments in establishing related guidelines and regulation.



Chapter 4: Implementation



This chapter provides an overview of key factors that support the safe, effective, and ethical implementation of AI technologies in local healthcare settings. These factors, which underpin aspects of the trustworthiness of AI technologies, were identified by interviewees for this research as key contributors to increasing confidence in AI among healthcare workers.

Confidence in the implementation of AI technologies can enable clinicians to assess appropriate levels of confidence in AI-derived information during clinical decision making, as detailed in **Chapter 5**.

4.1 Strategy and culture

Confidence that the right strategic decisions are being made about AI technologies in a culture that supports innovation and collaboration

The successful introduction and ongoing implementation of AI technologies in healthcare settings will depend, amongst other factors, on developing strong related business cases, maintaining effective relationships with industry innovators, and establishing organisational cultures conducive to innovation, collaboration, and public engagement.

These factors, which demand specific knowledge and skills, can influence how receptive workers will be to AI technologies and contribute to developing confidence in AI as a strategic and organisational asset.

The business case for adopting AI technologies

Interviewees for this research cautioned that some AI technologies are potentially being developed (including in proof of concept and trial stages) without a clear understanding of how they might affect healthcare provision. This underlines the importance of a strategic approach to the deployment of AI in healthcare settings, including through the development of value propositions and business cases.

For the interviewees, understanding the value, benefits, and risks of AI technologies (including in relation to patient outcomes, financial and human resourcing considerations, and alignment with related local and national strategies and priorities) are key to establishing strong business cases for deploying AI.

Despite available guidelines (including in the National Institute for Health and Care Excellence's evidence standards framework), the development of these business cases can be complicated by the minimal clinical evidence for most AI technologies (as discussed in **section 3.2**) and by workforce perceptions founded on limited knowledge and experiences of AI technologies.

Many interviewees for this research shared their concerns about the potential impact of AI technologies in their settings; for example, technologies that may lead to a higher number of patients being recalled and resulting in further costs for assessments and in unnecessary stress for their patients. Others wished that they were more cautious when they introduced AI in their settings by taking more time to debate and understand the related challenges.

On the flipside, interviewees voiced their hopes for using AI to address the major challenges in healthcare provision, including the increasing needs of an ageing population and the current backlog and waiting times in secondary care.

Relationships with industry innovators

Once AI technologies are procured, many are successfully embedded by establishing collaborative and sustainable relationships with industry innovators. These relationships can involve optimising and evaluating AI technologies.

Interviewees for this research noted that effective relationships between healthcare settings and industry innovators are built on shared values, and that they require a significant commitment of time and resources from both parties. Adept management of these relationships demands appropriate skills and resources, and shared understanding of the different perspectives within the AI ecosystem.

Interviewees noted that effective relationships can ensure buy-in and development of confidence in AI technologies among healthcare workers. However, they cautioned that the general lack of knowledge and experiences with AI technologies in most health and care settings limits their ability to critically appraise information provided by industry innovators, including during due diligence. These limitations in knowledge and experience may also lead some health and care settings to adopt industry-set parameters around the transparency and bias of AI technologies.

Interviewees noted that while some industry innovators support sites to establish a strategic approach to adopting AI technologies, others are not as proactive in their engagement or transparent about their products. Some interviewees spoke of 'hidden' costs involving AI technologies (including the costs of monitoring performance and reporting errors) that are not explicitly disclosed by some developers.

The importance of culture and leadership

Interviewees for this research highlighted that organisational cultures and leadership are key to the successful introduction and deployment of AI technologies. This insight is not surprising, and not exclusive to AI, as organisational cultures and leaders who support innovation and collaboration are crucial to the broader digital transformation of health and care settings.⁹

A focus on the digitalisation of health and care services is an important prerequisite to the adoption of AI technologies, as supported by NHS's What Good Looks Like framework (see also [section 3.3](#)).

Interviewees suggested a few key cultural and leadership features that can support these broader efforts, and increase confidence in AI, including:

- » developing AI technologies from the 'ground up' by involving multi-disciplinary teams (including clinicians, information technology and governance specialists, clinical domain experts and data scientists) and internal decision-makers early in discussions about their needs and implementation challenges. Interviewees for this research noted that multi-disciplinary teams tend to be the most successful structure for implementing AI
- » establishing senior leadership and clinical lead buy-in, and identifying and supporting internal champions for change

- » conducting early and ongoing engagement of patients and the public to inform AI development or co-design of technologies
- » focusing on ongoing learning and development of staff.

Interviewees for this research suggested that peer and expert endorsement and support can enhance confidence in AI technologies. This highlights the importance of developing and resourcing mechanisms to establish and encourage connections to share AI-related knowledge and experiences amongst peers and sites adopting AI technologies.

An example is the NHS AI Virtual Hub, an online platform for discussions, shared resources, and collaboration about AI technologies in health and care.⁶⁵ Other potential approaches can include developing and distributing case studies to highlight challenges and success stories (similar to existing NHS digital playbooks).⁶⁶

Key confidence insights

- » Confidence in AI as a strategic and organisational asset depends on developing strong business cases, maintaining effective relationships with industry innovators, and establishing organisational cultures conducive to innovation, collaboration, and public engagement.
- » Understanding the value, benefits and risks of AI technologies are key to establishing strong business cases for adopting AI.
- » Successful relationships between industry innovators and healthcare settings are built on shared values and require a significant commitment of time and resources from both parties.
- » Organisational cultures and leaders who support innovation and collaboration are key to the digital transformation of health and care settings, including the adoption of AI technologies.
- » Multi-disciplinary teams tend to be the most successful structure for implementing AI.
- » Developing and resourcing mechanisms to establish and encourage connections to share AI-related knowledge and experiences amongst peers and sites adopting AI technologies can support confidence in AI.



4.2 Technical implementation

Confidence that the implementation of AI technologies is supported by appropriate information technology infrastructures and data governance

The adoption of AI requires integration of these technologies with existing information technology (IT) infrastructures or the development of new IT infrastructures to support data storage, security, and information provision. It also requires the adaptation or development of information governance (IG) arrangements on data security, privacy, and clinical safety.

A key insight from the interviews conducted for this research is that IT and IG processes are a major factor in healthcare workers' confidence in AI technologies. Securing this level of confidence will require reaching internal agreements on the value of the data, the consent protocols, and the control, storage and use of the data.

Interviewees for this research noted that the broader technical challenges that relate to change and digital transformation in health and care settings complicate the adoption of AI technologies, including:

- » issues with hardware and software interoperability
- » time-consuming and impersonal processes for communication with IT support (for example, IT support being outsourced or located at separate sites)
- » unclear structures and responsibilities in IT and IG teams
- » miscommunications due to technical language and abbreviations.

Of these challenges, interoperability (compatibility and ease of integration with existing infrastructures) is particularly important. Interviewees suggested that interoperability is a frequent barrier to deploying new technologies in their settings, both from the perspective of users and industry innovators. For example, health and care settings can use various software and hardware infrastructures (including record systems, pathology systems, radiology systems and patient communication tools) that often require separate access details. AI technologies that operate as separate applications would likely frustrate workers and limit their uptake.

More importantly, a site's ability to adopt AI technologies can rely on its capacity and resources to support related IT and IG arrangements and other requirements (including commercial agreements and Data Protection Impact Assessments).

Discrepancies between clinical departments at each site can also influence the ability to adopt AI; for example, some departments maintain separate IT infrastructures to other departments (like cardiology and radiology) that impact how data is collated and stored. These discrepancies can result in inconsistent

and incomplete data sets. Developing and maintaining comprehensive and representative data at a scale required by AI technologies remains a significant limitation of the healthcare system.

Interviewees cautioned that, although infrastructure and data-related challenges are not specific to AI, they must be resolved to achieve broad deployment of, and support confidence in, these technologies in health and care settings.

Interviewees suggested that the adoption of AI technologies should be integrated into broader digital transformation systems and involve coordinated multi-disciplinary teams across clinical, technical, and administrative roles, potentially across different settings (as noted in **section 4.1** and detailed in the second report). One interviewee suggested streamlining approvals for related infrastructures through centralised networks or clusters (for example, through the Integrated Stroke Networks for radiology-related technologies).

This connection between the adoption of AI technologies and broader efforts to implement change and digital transformation in healthcare settings confirms the importance of incorporating change management skills and enhancing digital literacy amongst the workforce. HEE is currently undergoing an extensive Digital Readiness programme to enable staff to identify their digital readiness and meet their training needs.

Interviewees for this research noted that the COVID-19 pandemic has assisted in changing attitudes towards developing new infrastructures and being open to changing the status quo in health and care settings. Many healthcare workers have been 'forced' to reconsider existing IT and IG infrastructures to address the growing backlog of required services, essentially embracing the importance and urgency of digital health and digital transformations.

The development and dissemination of AI-related resources and guidelines could also assist health and care settings to overcome potential challenges related to their IT and IG systems. Existing resources, like the NHS's Interoperability Toolkit,⁶⁷ can provide a blueprint for AI-specific guidelines. Interviewees for this research spoke also of the need for standardised information on related terminologies (including on the efficiency, safety, and performance of different AI technologies) and common approaches to IT, IG, data security, and patient privacy.

Sharing of knowledge and experiences relating to the challenges of implementing AI in healthcare settings (for example through creating communities of practice) could also be beneficial (as noted also in **section 4.1**). A few industry innovators have developed their own initiatives by providing 'start-up packs' to guide the technical deployment of their technologies within specific settings, and by coordinating peer support groups amongst their deployment sites.



Key confidence insights

- » Establishing and agreeing on related IG and IT arrangements are instrumental to healthcare workers' confidence in using AI technologies.
- » Ideally, the adoption of AI technologies should be integrated in broader digital transformation systems and should involve coordinated multi-disciplinary teams across clinical, technical, and administrative roles.
- » The development and dissemination of AI-related resources and guidelines can assist health and care settings to overcome potential challenges related to their IT and IG systems.

4.3 Local validation

Confidence that AI technologies are having the right impact locally

Section 3.2 provides a detailed discussion on the various evaluation and validation approaches for AI technologies, including local validation.

Local validation of AI technologies may be needed to ensure that published data on the performance of AI technologies are reproducible in the local context. Such validations will vary depending on the technology and how the technology will be implemented, and may also involve distinct local settings or clusters: from individual practices to Integrated Care Systems.

There are many unknowns and potential risks involved in 'translating' AI technologies from controlled development and validation settings to complex and highly individual real-world settings. These risks can relate to the ability of settings to understand the suitability and performance of the AI technologies locally (including in relation to local populations, practices, hardware, and data pipelines), to maintain the ongoing rigour of that performance, and to minimise any unfair impact on or harm to their patients.

Interviewees for this research noted that being unable to assess whether an AI technology is suitable to their local settings and populations may contribute to unease and hesitancy to adopt AI technologies within workers. The 'blind' acceptance of validations conducted in different settings and populations is an additional risk.

However, interviewees were cautious as to whether each health setting can sustain the resources required for the local validation and ongoing maintenance of AI technologies, including the capacity of clinicians to participate in these processes. They noted that clinicians may hesitate to adopt any AI technologies that expand their workload unless mandated to do so. They may not welcome additional responsibilities for acting as 'gatekeepers' of AI, especially without sufficient specialist knowledge and training. Interviewees suggested that this responsibility could be undertaken by local specialists (for example clinical scientists) or centralised entities (as discussed in the **second report**).

Interviewees suggested that conducting research related to AI (including in relation to evaluation and validation) is an excellent training opportunity for internal teams to identify risks and understand the benefits and value of AI technologies.

However, they noted that these opportunities are typically only available to staff at larger healthcare centres. Creating opportunities for staff from smaller organisations to be seconded onto exemplar projects could be encouraged to support this model of training.

Furthermore, hands-on experience under the guidance of expert peers could be encouraged alongside educational programmes and materials. Centres with expertise in these areas could disseminate their knowledge and support colleagues in organisations with less experience (as discussed also in **sections 4.1** and **4.2**).

Key confidence insights

- » Health and care settings will need to understand the suitability and performance of AI technologies locally (including in relation to local populations, practices, hardware and data pipelines), maintain the ongoing rigour of that performance (post-market surveillance), and minimise any unfair impact or harm on their patients.
- » Providing universal opportunities for staff to engage with experts and AI-related research and validation projects could enhance their confidence in AI.



4.4 Systems impact

Confidence that AI technologies are being integrated in clinical workflows and pathways in a safe, efficient, and ethical manner

Interviewees for this research noted that confidence in AI technologies will depend on perceptions of their safe and efficient integration into clinical workflows and pathways.

Other research has found that AI technologies need to work quickly, reliably and effectively to instil confidence in the healthcare workforce,⁹ and this was reemphasised by clinician interviewees for this research. They noted that front line healthcare workers are often frustrated by unreliable hardware and software that impact their ability to deliver good quality care. They perceived NHS healthcare software such as electronic health records (EHRs) to be difficult to use, requiring extensive training and support, and with limited capacity for customisation to meet user needs.

Interviewees expected new technologies to improve on these legacy systems by being user-friendly, intuitive, and where possible, customisable. For example, the use of AI technologies should preferably not require logging into separate systems, and if appropriate, AI-derived information should be stored as part of the patient record, not separately. Ideally, AI technologies will streamline existing workflows, and reduce complexity by processing data automatically. Seamless integration will also enable robust working practices for the ongoing monitoring, evaluation and audit of AI technologies, making good practice easier to achieve and building system-level confidence.

Interviewees noted that health and care settings may need to review and revise their existing systems through new ways of thinking about clinical practices, patients, and their care. Expanding their capabilities to change and adapt, including in informatics and data collection will also be important. Broader efforts for digital transformation and change management can support such transitions as described in **sections 4.1** and **4.2**.

Interviewees cautioned that AI technologies that entail higher clinical consequences relating to patient triage, diagnosis or care will require appropriate measures to secure patient safety, and to clarify steps for reporting adverse effects in case of system failure. These could include error reporting pathways, effective use of national reporting (for example through regulatory requirements), and fallback workflows for system failures or unsuitable user cases. Development of protocols to clarify related actions and ensure human oversight will be essential.

Post-market surveillance of medical device technologies is currently managed through the MHRA's Yellow Card system. The NHS AI Lab is currently supporting the re-design of this system using data-driven technologies to better identify and track trends with incidents of adverse performance.

In addition to efficiency and safety, an ethical approach to AI is essential to achieving confidence in these technologies. Interviewees for this research highlighted that an ethical approach to implementing AI should include, at a minimum, the principles of fairness, transparency, and accountability, and aim to ensure equitable benefits across patient groups.

Finally, the way in which AI is integrated into clinical workflows and pathways may impact clinical decisions. As explored in greater detail in **Chapter 5**, research suggests that clinicians who perceive themselves to be domain experts will typically view and use AI-derived information differently to non-specialists, especially in time-pressured decision making.⁶⁸ This suggests that testing the real-world impact of implementing AI, including the timing and manner of data presentation may be necessary to allow clinicians to correctly assess and use AI-derived information.

Key confidence insights

- » Healthcare workers will be more confident in AI technologies that are safely, efficiently, and ethically integrated in clinical workflows and pathways.
- » Ideally, AI technologies should streamline existing workflows and be seamlessly integrated to improve their adoption.
- » Clear pathways should be established for reporting safety events with AI technologies.
- » An ethical approach to AI will be essential to achieving confidence in AI. At a minimum, this should include the principles of fairness, transparency, and accountability, and aim to ensure equitable benefits across patient groups.
- » The way in which AI is integrated into clinical workflows and pathways may impact clinical decisions and should be considered during its design process.





Chapter 5: Clinical Use



This chapter discusses the challenges of incorporating AI technologies into clinical reasoning and decision making (CRDM), and determining the **appropriate level of confidence** a clinician can place in AI-derived information (the output provided by an AI model to a clinician) for a case-specific clinical decision.

Many AI technologies used within health and care settings do not directly affect CRDM (for example technologies that support workflow optimisation and scheduling like appointment booking tools). This chapter is not relevant to these applications of AI, for which confidence is built through trustworthiness, based on the factors described in **Chapters 3** and **4**.

In other parts of the report, the term AI 'user' encompasses clinical and non-clinical (for example, administrative) users of any AI product used in healthcare. In this chapter, a 'user' is specifically a clinician who uses AI technologies to assist with, enhance or perform CRDM that will directly affect patient care. This may include screening, health monitoring, diagnostics, prognostics, treatment stratification, design, optimisation, response monitoring or any other clinical aspect of a patients' care pathway.

The chapter provides an overview of key aspects of CRDM, addresses the factors affecting confidence in AI-derived information at the point of CRDM, and discusses the challenges of enabling clinicians to know when they have appropriate confidence in AI-derived information.

As described in **section 2.2**, appropriate confidence in AI-derived information is also supported by confidence derived from the trustworthiness of AI technologies and their implementation, through the factors presented in **Chapters 3** and **4**.

5.1 Aspects of clinical reasoning and decision making (CRDM)

5.1.1 Assessing confidence in clinical information at the individual patient level

Clinicians interviewed for this research discussed the complexities of CRDM processes and the assessments of various types of information that are involved.

They described conventional, human CRDM as a complex and nuanced process, where clinician experience, intuition, expertise, and biases interact, often subconsciously. Clinicians are experts in assessing the appropriate level of confidence to have in various forms of information feeding into CRDM, from patient histories to test results, imaging, and reports from professional colleagues.

For the clinician interviewees, effective CRDM requires them to make value judgments about the significance and trustworthiness of information derived from sources of either unknown reliability (for example, patient history) or reliability that has been demonstrated at a cohort or population level (for example, laboratory test results). They combine that information, which can be potentially contradictory, to make an optimal decision with each patient.

The individual nature of any clinical decision places the responsibility on the clinician to know how to weigh up, or indeed when to disregard, certain information. Clinicians can base their decisions on a complex synthesis of knowledge, professional experience, patient history, demographic factors, test results, imaging, expert opinion, patient preference and intuition.

Through assimilation of all this information and given their expectations as to the most likely clinical scenario, clinicians make implicit or explicit probabilistic estimates, determine a course of action with the patient as appropriate, and act accordingly. Interviewees noted that clinicians learn these skills over an extended period through training and experience.

This context suggests that the CRDM process, while guided by best-practice, guidelines and published literature, can also be highly individual and contextual.

When faced with uncertainty, clinicians may access peer opinion, discuss conflicting views and seek the opinion of a multidisciplinary team if required. Some clinical decisions (for example, in emergency medicine) are made rapidly under time pressure, while others are made in a very considered way and without such urgency, potentially in consultation with other experts or considering recent academic literature.

Interviewees noted that these nuanced and experience-driven CRDM processes are susceptible to cognitive biases, which can be exacerbated by the unreliability of human intuitions about probability, and the perceived trustworthiness of the provided information. Confirmation bias and automation bias (as defined and discussed in **section 3.3.1**) are common challenges in existing CRDM situations, and it is important to consider these carefully when new and unfamiliar tools or information sources, such as AI decision-support tools, are introduced into CRDM.

Interviewees concluded that AI technologies provide both risks and opportunities in this context: cognitive biases may either be reinforced, or mitigated, by the inclusion of AI-derived information in the CRDM process. This suggests the importance of clinicians understanding how their current decision making process could be influenced by the introduction of AI technologies, particularly in the case of conflict between their own intuition or opinion and the information or recommendation provided by an AI system.

5.1.2 Confidence during AI-assisted CRDM

AI technologies that are used to support clinical reasoning and decision making can be referred to as AI-assisted CRDM or clinical decision support systems.

Incorporating AI-derived information into CRDM has enormous potential to improve consistency and quality of clinical decisions, increase efficiency, and benefit patients.⁶⁹ In AI-assisted CRDM, the clinician retains ultimate responsibility for the decision made, so it has been suggested that clinical 'reasoning support', as opposed to 'decision support' is more appropriate terminology.⁷⁰

As highlighted by this research's interviewees, clinicians who use AI-derived information during CRDM will need to understand the nature and context of this information to assess whether it warrants low or high confidence.

Interviewees suggested that awareness that most AI technologies involve statistical methods and probabilities to make predictions is a fundamental starting point for anyone using AI in CRDM. For example, AI models can be trained on existing data to predict the most likely diagnosis or the treatment strategy with the best chance of success. This approach relies on the assumption that the present case and situation are similar enough to those used to train the AI system.

For this research's interviewees, AI-derived information should be perceived as a prediction or estimate of the most likely diagnosis or optimal strategy and should be considered to have a degree of uncertainty associated with it, in the same way an external opinion might be assessed.

The individual reliability of AI-derived information will be a function of the data used to train the model, the training process itself, and the characteristics of the particular patient for which a prediction is being made.

Interviewees observed that clinicians are used to these estimates of reliability for non-AI information sources. For example, clinicians are aware that laboratory results can on occasion be incorrect, and may be aware of their statistical performance at a population level. A known example involves the common prostate-specific antigen (PSA) test for prostate cancer. It has a 75 per cent false positive rate and a 15 per cent false negative rate.^{71,72} Clinicians are used to interpreting results such as PSA in clinical decision making and counselling patients about the potential unreliability of such tests.

AI-derived information is potentially different to test results or other quantitative measurements, where the likelihood of error remains largely constant across broad categories of patients, and are known to vary predictably based on factors like patient demographics. In the case of AI, the aspects of the data impacting the accuracy of the prediction may be more complex and often unknown,⁷³ making it harder for clinicians to assess how confident they should be in a specific AI prediction (as discussed further in **5.2.3**).

'Brittleness' (the tendency for performance to fall off rapidly at the boundaries of the AI algorithm's scope)⁷⁴ in the clinical use of AI is a particular challenge to determining appropriate levels of confidence during CRDM.⁷⁵ It suggests that, when applying population-level evidence and performance metrics to AI predictions concerning individual patients, clinicians will need to be cautious and retain a critical eye for unexpected, contradictory or implausible predictions (as discussed also in **Box 5** and **Box 6**).

Further, failure cases for an AI's performance may or may not be similar to those where human performance is low,⁶⁸ making identification of these error cases particularly challenging. This is an additional reason for clinicians to retain a critical eye when dealing with AI-derived information for CRDM.

Therefore, even when confidence in an AI technology (as derived from the factors presented in **Chapters 3** and **4**) can be high, clinicians should still be encouraged to question predictions that appear to go against their clinical intuition or other evidence, remembering that appropriate confidence in that specific case may need to be low.⁷⁶

It is important, however, to note that AI technologies can potentially find correlations in the data of an individual patient that human CRDM would not account for.⁷⁷ In such cases, it is possible that an AI-derived prediction or recommendation that is contrary to a clinician's intuition may be correct and should be considered seriously rather than dismissed out-of-hand.

These considerations suggest that educating clinicians to retain a degree of scepticism in AI-derived information, while not losing confidence in the overall performance of the AI technology, is an important aspect of practising CRDM with AI technologies. If done well, AI-assisted CRDM has been shown to have the potential to outperform both human and automated approaches.⁷⁸

Key confidence insights

- » Clinical Reasoning and Decision Making (CRDM) is a complex, nuanced process, learned through lengthy education and professional experience. It relies on making value-judgements about information from a range of sources.
- » Appropriate confidence in AI-derived information should be assessed for each patient and each AI-assisted clinical decision.
- » Clinicians who use AI-derived information during CRDM will need to understand the nature and context of this information to assess whether it warrants low or high confidence.
- » AI-derived information should be perceived as a prediction or estimate of the most likely diagnosis or optimal strategy and should be considered to have a degree of uncertainty associated with it, in the same way an external opinion might be assessed.
- » Clinicians need to understand how their current decision making process could be affected by AI-derived information and understand the importance of retaining a critical eye, to detect potential AI failure cases.
- » Education and training will be key to developing appropriate levels of confidence during CRDM.

5.2 Factors affecting confidence in AI during CRDM

This section outlines factors that can affect a clinician's confidence in AI during CRDM. These involve the complex and nuanced interactions between the clinician's personal attitudes and experiences, the clinical context, and the implications of the various characteristics of AI models.

5.2.1 Impact of individual attitudes and experiences

Interviewees for this research noted that personal experiences and attitudes to innovation and AI technologies can strongly affect a clinician's confidence in using AI technologies.

These factors correlate well with those identified in previous research concerning trust and AI technologies: ^{4,34,79}

- » **General digital literacy.** Digital literacy is a key enabler in the adoption of new technologies. Literacy can vary across age groups, professional groups and places of work, with younger professionals in larger centres more likely to have good digital literacy in general.⁵
- » **Familiarity with technology and computer systems in the workplace, and past experiences with AI or other innovations.** Those more familiar with AI are generally more positive about its use in healthcare. However, a little familiarity can be dangerous, if it leads to an unquestioning preference for AI-derived information. ^{2,34,79}
- » **Personal attitudes towards technological innovations at work, including the perceived risks to an individual's role through automation.** Early computerised clinical tools have often performed poorly and generated more work for clinicians, leading to technology scepticism amongst some more experienced clinicians. Hesitance amongst healthcare workers may also be driven by media scepticism, reports of biased AI and data privacy concerns, as well as fears of losing their roles and responsibilities.⁷⁹
- » In addition, several interviewees noted that **ethical concerns** around fairness and bias in the development and use of AI in healthcare can potentially affect confidence in these technologies.

Research has shown that clinicians are more likely to trust AI-derived information that does not relate to their **area of expertise**.⁸⁰ Conversely, experts (specialists in their area of care) tend to be more sceptical, questioning AI-derived information more than generalists or more junior colleagues.⁶⁸

Some experts interviewed for this research suggested that they can perceive AI-derived suggestions as 'equivalent at best' to their judgement, and potentially inferior in many situations. They may also have different perspectives of the impact of different decisions, compared to non-specialists. For example, they may view the importance of the AI-derived information differently when judging the performance of the AI technology based on clinically relevant endpoints (e.g. impact on treatment or outcome), rather than numerical accuracy of the AI prediction itself.

However, it is important to note that the performance of human experts measured in trials or evaluation settings is rarely maintained with consistency in routine practice, due to quantity of work, fatigue and external distractions.⁸¹ As AI technologies do not suffer these human limitations, they may offer a benefit even in situations where experts currently perform these tasks.

Importantly, in many situations, not all clinical decisions are made by experts, and certainly not in the first instance. Supporting decision making in this context with AI-derived information could assist non-specialist clinicians in pressured scenarios and improve the quality and consistency of patient care. However, this may lead to non-specialists no longer having the opportunities to develop specialist skills, potentially resulting in de-skilling and a future skills shortage.

Interviewees for this research cautioned that non-specialist, or junior clinicians who need to make decisions on treatments and conditions that they have limited experience with, may tend to willingly accept and favour the support provided by an AI technology. This leads to the potential for uncritical adoption of AI predictions and resulting bias, a finding in agreement with previously published research.⁸²

These variations in clinician experience and attitudes suggest several challenges:

- » experts may be sceptical of the benefits of AI to their work, leading to missed opportunities to improve consistency and quality in CRDM.
- » less experienced clinicians are potentially susceptible to inappropriate high levels of confidence in AI-derived information and may require education to be able to critically appraise this information.
- » non-expert groups may come to rely on technologies that minimise the stress and complexity of their decision making, in place of developing their skills and ongoing learning, which will lead to de-skilling parts of the workforce.⁸³

5.2.2 Impact of clinical context

The clinical context in which AI is used can influence confidence in the technologies and in the derived information, including in relation to the levels of clinical risk and human involvement.

Level of clinical risk

The level of involved clinical risk can influence confidence in AI technologies. Interviewees for this research predicted that healthcare workers will be more confident in using AI technologies that have lower clinical consequences, such as tools that prioritise certain cases for urgent review, but do not alter the clinical input that any patient ultimately receives. This is consistent with other research demonstrating that clinicians were more positive about using AI for administrative tasks rather than clinical tasks.⁸⁴

This propensity suggests that AI technologies with a higher risk of clinical consequences may be adopted more cautiously, and that healthcare workers may expect a higher threshold for evidence of the safety and efficacy of these

technologies. Interviewees cautioned that, in such high-risk instances, it is important to take a holistic view of the clinical situation and consider an AI prediction as one contributor amongst many to clinical decision making.

Level of clinician oversight

Feedback from healthcare workers and the public in other research suggest that AI-assisted CRDM technologies are more acceptable than technologies that act autonomously (referred to as autonomous AI), without human involvement in each decision. Tools that perform triage and result in patients not receiving review by a human clinician are examples of autonomous AI and should therefore be considered as 'high-risk'.

These views are associated with a preference for a 'human touch' in the provision of healthcare services, both in relation to decision making and communication with patients.^{85,86}

The preference for AI-assisted CRDM rather than autonomous AI may have contributed, along with commercial considerations, to a trend observed through the interviews conducted for this research: that industry innovators are following a 'human oversight' approach when developing AI technologies. For example, AI technologies used for screening tasks are designed to be deployed alongside human graders rather than to replace human graders.

The reluctance to develop and deploy fully autonomous AI may also be due in part to the limited available evidence supporting their clinical efficacy. Unresolved governance issues including regulation and liability for autonomous AI technologies may add to this uncertainty, making organisations and individuals understandably more risk-averse in this context. These external factors are further discussed in **sections 3.1** and **3.4**.

5.2.3 Impact of AI model design

Interviewees for this research described various characteristics of AI models that can influence confidence. These are described in detail in this section.

Interviewees noted that most healthcare workers do not have sufficient knowledge and understanding to ask meaningful questions about the features presented in this section. This suggests that part of the educational challenge around AI for CRDM is to equip clinicians to ask these questions, and to understand the significance of the answers across various clinical scenarios.

The nature and presentation of AI-derived information for CRDM

There are certain characteristics of AI-derived information that can make an assessment of appropriate confidence during CRDM challenging.

All AI systems use probabilistic methods to make predictions, regardless of whether they present that prediction categorically (for example, diagnosis A vs. diagnosis B) or as a probability (for example, by providing the predicted probability


of a particular diagnosis). While some conventional types of clinical information are also understood using probability (for example, a test result that has an associated sensitivity and specificity), AI technologies involve more complex and subtle statistical analyses that are harder to interpret and communicate effectively (described as the 'black box' problem of AI). This can lead to a reduced ability to understand the implications of a probabilistic output, and complicate determining the appropriate confidence in the information.⁸⁷

AI technologies can be designed to include certain features with the potential to demystify and provide some degree of confidence as to how reliably they make predictions. These include features like uncertainty quantification, outlier detection and clarifying input data, as described in **Box 5**.

Interviewees noted that the way AI-derived information is presented can influence a clinician's confidence in that information. This suggests, as also supported by literature,^{88,89,90} that careful user interface design, and the way in which AI-derived information is presented, can support assessment of appropriate confidence during CRDM.

AI-assisted CRDM can present a categorically different type of information to other forms of input, such as a numerical test result or patient history.

An AI technology may present to the clinician a 'risk score', suggested diagnosis or course of action, which can be perceived as synthesised information, similar to an opinion or the end result of the CRDM process rather than a conventional piece of factual input information. A clinician can associate a test result with a diagnosis or risk level, weighing other patient factors in the process, whereas the AI-derived information may already incorporate some or all this information. This can make assessing confidence in the AI-derived information more challenging, unless it is very clear what information has been considered by the algorithm, and what relative weighting it has been given.⁹¹

 *Example: A prostate cancer risk stratification AI, based on patient demographics and imaging, predicts a high risk that contradicts the clinician's intuition. The patient is on an unusual medication that may affect the appearance of their prostate on imaging. The clinician, therefore, has reason to question the AI in this case and rightly has lower confidence in the AI risk score. They should default to making their own assessment of the imaging and associated patient factors.*

Compared to those providing only categorical predictions, AI technologies that provide uncertainty quantifications can support confidence assessment for CRDM by giving users a sense of the algorithms' internal certainty level. For example, if an algorithm is 80 per cent certain, then it should produce the correct prediction 80 per cent of the time. This is known as the probabilities being 'well-calibrated'.

Box 5. Features that can improve confidence in how AI algorithms make predictions

Uncertainty quantification

AI technologies may quantify indicators for uncertainty in their predictions. This can be shown in various ways, including confidence statements, scales, scores or even raw probabilities. These different ways of presenting uncertainty quantification can influence how users perceive and act on that information.

Q *Clinical example: An AI product is designed to diagnose pneumonia on a chest X-Ray. When it provides the report, it states how confident it is that there is pneumonia on the X-ray. The user may see the following output- 'diagnosis: pneumonia-high probability, whereas another could receive the following output- 'diagnosis: pneumonia- low probability'.*

The value of showing uncertainty quantification and the form in which it is displayed will vary depending on the task and the clinical setting. More research is needed to investigate how displaying uncertainty quantification can impact clinical decision making, and product-specific guidance may be required to assist clinicians to interpret this information.

Data integrity

It is possible for an AI to be incorrect but present high certainty in this incorrect output. This is likely to occur if the input data is further from the centre of the distribution of the training data, a situation known as 'algorithmic brittle. Cases that are towards the limits of the distribution, known as edge cases, may occur because of continuously variable patient factors (for example, age, size, height or weight).

Cases that are extremely dissimilar (several standard deviations) from the training distribution are known as outliers and are at high risk of erroneous predictions. Technical factors (for example, different radiology equipment used to take images, image brightness, resolution etc), or categorical patient features (gender, ethnicity, surgery, co-morbidity) can commonly cause outliers. In the extreme case, this type of data could be considered invalid input for the algorithm.

Outlier detection can mitigate against these situations, by estimating the similarity of the clinical case to the training data distribution and alerting the user to dissimilar cases. AI technologies can be designed to include outlier detection, increasing user confidence by ensuring that the AI technology will be able to highlight cases that may require additional clinical review and consideration.

Box 5. (cont.)

Q *Clinical example: An AI product has been designed to detect lung tumours in chest X-rays. A patient with a lobectomy is imaged. No lobectomy X-rays were included in the training data. The product identifies this image as unusual compared to its training data, and alerts the user that it should be considered less reliable than normal for this case.*

Scope of input data

Knowledge of the input data and the factors being considered by the AI's computational methods can enhance user confidence. Without this information, workers may be uncertain about the value of AI-derived information for particular patients.

Q *Clinical example: An AI product is designed to predict the risk of skin cancer based on images of skin lesions. A primary care physician tests a lesion on a patient who has significant risk factors for skin cancer. The AI returns a 'benign' output. The confidence of the clinician trusting this output may be influenced by knowing what factors the AI has used to make the decision. For example, if the clinician doesn't know whether the AI has taken the patient's risk factors into account, they may be more likely to ignore the AI output and refer this lesion for review by secondary care; however, if they know the AI has considered these, they may be more likely to reassure the patient about the 'benign' output.*

However, inaccurate calibration (for example, an algorithm estimating 80 per cent certainty, but when tested found to produce the correct prediction only 70 per cent of the time) has been shown to be common in AI, particularly with deep-neural networks.⁹² This is likely to lead to inappropriate levels of confidence, due to an impression that the probabilities will accurately indicate uncertainty for each case, when they may not. Therefore, before uncertainty quantification is used to support the assessment of appropriate confidence for AI-assisted CRDM, good calibration of these probabilities should be demonstrated in a validation cohort.

The impact of miscalibration in AI models has been shown to be exacerbated when non-expert users were involved in shared decision making, as the humans were unable to bring in enough unique knowledge to identify and correct the AI's errors.⁹³

Conversely, presenting AI predictions only categorically (for example, without uncertainty quantification) may lead to inappropriate assessments of this information. Interviewees were divided on whether categorical predictions were

more likely to cause inappropriately high confidence (in the case of the predictions being taken at face value) or inappropriately low confidence (due to a perceived lack of nuance). They hypothesised that individual attitudes towards AI may influence how categorical predictions may be perceived and suggested that decisions around uncertainty quantification and clarity of information must be balanced to suit the use case. Some interviewees noted that they would always value additional contextual information, whereas others preferred a decisive, categorical recommendation over a probability, preferring the AI to simply refuse to offer an opinion at all, if internal uncertainty was too high. It was suggested that clinicians under time pressure, and especially when at the boundaries of their expertise, may prefer a 'decisive' presentation of information.

Interviewees concluded that the ideal approach to presenting AI-derived information depends on the clinical context, and should be considered carefully by algorithm and software designers, as well as implementors in health and care settings. Users should also be involved in the AI's evaluation and implementation processes to determine the best approach for the presentation of the AI-derived information. Additional research is needed to fully understand and optimise the nature and timing of presentation of AI-derived information for CRDM across the range of relevant clinical settings and scenarios.

Further, interviewees suggested that AI-derived information may best be considered as if provided by an external specialist (for example a radiologist or pathologist), who will also typically have access to some but not all of the relevant clinical information and will have a different locus of expertise to the clinician themselves. In conventional CRDM processes, clinicians understand when they have additional information that influences the diagnosis or treatment choice, above that which the reporting specialist had access to. This same approach could be followed to assess the appropriate level of confidence in AI-derived information.

Explainability

Research suggests that transparency in how AI computes and delivers an output, and the possibility of proving a 'human-like explanation' for the prediction (referred to as explainability) encourages higher levels of confidence in the technology.⁹⁴ This conclusion was supported by interviewees who highlighted a desire for explainability when designing and procuring AI technologies.

Some AI systems may be able to explicitly describe the importance being given to inputs or features of the data, enabling a direct explanation of their relative importance. However, many AI technologies, particularly those that use deep-learning, can be described as 'black box' as their algorithms do not inherently explain their 'reasoning'. It has been argued that current machine-learning methods do not 'reason' like humans,⁹⁵ but rather identify patterns in data and correlate them with previous human decisions or outcome labels, 'learning' the best predictive model for a human decision, but without the 'reasoning' element. In machine learning, there is no guarantee that the features identified by the AI are the same ones the humans use in their reasoning processes, or that features


that are shared with human observers are selected for the same reasons. There is a possibility that the AI may rely on features that are not appropriate to the task (such as laterality markers, or equipment characteristics in medical images),⁹⁶ or detect features that are of completely unknown relevance to clinicians.

Explainable AI (XAI) involves approaches that attempt to 'shine light' into an AI 'black-box'. XAI has been heralded as providing an intuitive picture of the features of the data and 'reasoning' that drive AI predictions, similar to asking a clinician what reason they have for a particular decision.⁹⁷

However, emerging evidence⁹⁷⁻¹⁰⁰ suggests that current XAI approaches to deep-learning models can be misleading for assessing individual predictions. The 'explanations' provided can bear little relevance to finding out if the individual prediction was made for solid reasons.

XAI methods can only highlight salient parts of the input data and do not give clarity on the clinical relevance of the features to the individual prediction, or the ways in which they have been combined. XAI may therefore not correlate well with clinicians' views of what an 'explanation' should be.⁹⁸ There is some recent evidence⁹⁹ that XAI techniques for image classification can actually be independent of the trained variables in the AI model (what the model has learned) and the data-label relationship (what the model tries to predict), instead merely highlighting 'interesting' parts of an image, such as edges and texturally rich areas which contain the most information.

Further, research has found that labelling a product as XAI may make its predictions appear inappropriately trustworthy.¹⁰⁰ XAI can potentially provide confidence that an algorithm is generally looking at the clinically relevant parts of an image (for example, lung boundaries for pneumothorax detection). However, XAI is not useful at an individual case level to determine the reliability of an AI prediction.^{100,101}

 *Example: A chest X-ray diagnostic algorithm for pneumothorax with Grad-CAM XAI provides associated saliency maps, which highlight certain areas of the lung that are 'important' to the prediction. However, the same regions are identified for most patients, irrespective of whether pneumothorax is present or where it is located in the image. The saliency map is identifying the image regions that are important at a model level, but not for the individual patient in question.*

Some progress has been made¹⁰² in creating AI for chest X-ray screening that can provide 'human-like' explanations. The aspects of an image that contribute to a prediction of malignancy are highlighted and described in terms of their similarity to previous known-malignant examples. This process mimics how a human observer might explain their reasoning. This approach fundamentally differs from the common approach of adding post hoc explainability to a deep-learning AI model through saliency maps. The approach required a complete re-imagining and re-engineering of the AI architecture and training, to ensure it inherently makes predictions using 'human-like reasoning'.

Some interviewees for this research were more likely to be sceptical of a 'black-box' AI than an XAI system. However, given the concerns around the suitability of XAI for individual case-level analysis, it may be inappropriate to consider XAI as a panacea for clinical confidence assessment at the individual prediction level. For individual predictions, it may be safer to assume that no 'explanation' is possible, and the system is treated as a 'black-box'. This approach would mitigate the risk of inappropriate high level of confidence resulting from the 'veneer of explainability' provided by current XAI methods.¹⁰⁰

Bias in AI models

Interviewees for this research were particularly concerned about the potential for bias in AI technologies, and how bias affects confidence in the performance of these technologies for different patients.

AI technologies are trained on data that reflect human decisions, which leads to the possibility of reinforcing or perpetuating biased, unfair or unethical tendencies amongst human operators.¹⁰³ For example, AI systems that aim to triage patients or prioritise referrals, based on historical human decisions need to be carefully designed and validated to mitigate biases involving underlying access or communication challenges in certain patient demographics or cohorts.

Bias in AI algorithms has various sources,¹⁰⁴ from bias in the demographic profile of the training data set, to the quality of the labelled data, the technical design of the model, and the team designing the algorithm.¹⁰⁵ **Box 6** provides some examples of these types of bias in AI models.

Potential biases in AI models and their implications for patient safety and fairness are a key reason that external validation is critical to building confidence and safety in AI for healthcare (as discussed in **section 3.2**). The concept of bias towards or against certain sub-populations in the clinical environment is closely related to the 'generalisability' of an algorithm,⁷³ but it is important to dig deeper than the whole-cohort evaluation to determine whether certain groups may be disadvantaged or experience lower algorithm performance.

Creating well balanced, labelled data sets of sufficient size and quality can be extremely difficult (due to data unavailability) in healthcare, and the potential to embed bias into AI models is therefore high. For example, data collection can be imbalanced by the inclusion of rare diseases that are overrepresented in retrospective data sets.

Box 6. Examples of AI biases

Data bias

Data bias is the root of much algorithmic bias in AI and can take several related forms.¹⁰⁶ The most obvious is a simple imbalance of populations, or absence of a particular demographic, in a data set, leading to a lower performance for the minority group. However, the problems may be more subtle and complex.

In an example¹⁰⁷ of racial bias in a widely used risk-scoring algorithm, Black patients assigned the same level of risk by the algorithm were found to be sicker than White patients, reducing the number of Black patients identified for extra care by more than half. Bias occurred because the algorithm used health costs as a proxy for health needs. As less money is spent on Black patients who have the same level of need, due to health inequalities, the algorithm thus falsely concluded that Black patients are healthier than equally sick White patients and require less care. This is an example of poor AI design reinforcing and concretising health inequalities that are pre-existing in the current system.

A recent study¹⁰⁸ examined the performance of a convolutional neural network classifier for skin lesion identification in a sub-Saharan African population. The researchers found overall very poor performance (accuracy of 17%) and determined this to be data bias driven in two ways. Firstly, the number of dark skin type images included in training was low and secondly the disease burden in the African population was significantly different to the western Caucasian population used in model development. Hence, many conditions present in that location were underrepresented or absent from the model.

Racial inequality is not the only risk deriving from data bias. Care must be taken around inequalities related to gender, age, (dis)ability, socio-economics, faith, geography, and any other demographic factors that could potentially disadvantage certain patients.

Some data sets used for healthcare AI model development may be completely inappropriate, due to confounding factors and lack of appropriate data collections methods, leading to serious generalisability concerns. A recent meta-analysis of COVID-19 CXR detection algorithms¹⁰⁹ determined that many had been trained on publicly available 'Frankenstein data sets' where the positives were all from one healthcare site and the negatives from another. Most algorithms were found to be learning to identify positive cases based on non-clinical 'shortcuts' (e.g. style of laterality marker or site-specific scanner characteristics) rather than clinical signs of COVID pneumonia, leading to an almost complete lack of clinical usability, when taken outside the development data set. This situation can hardly be described as 'bias', but less severe examples could easily lead to uneven performance across sub-populations and hence strong biases.

Box 6. (cont.)

A more subtle, but related form of data bias is in the application of well-validated AI technologies to a clinical situation that lies towards the edge of the algorithm's experience. In such a case, the input data (patient cohort) is more likely to contain cases that are 'outliers' to the algorithm. This can lead to a dramatic fall in the performance of the AI technology, even following extensive external and local validation.¹¹⁰ In this context, a generally trustworthy AI algorithm may exhibit 'brittleness', by making untrustworthy predictions in some specific cases (the subset of patients who lie towards the edge or outside the training data distribution). This is a challenge that requires approaches such as outlier detection from the product point of view, as well as clinical vigilance from a user perspective (as discussed also in **Box 5**).

Team bias

The makeup of AI development teams can impact issues of fairness and bias in AI. The lack of gender and ethnic diversity in AI teams can contribute to perpetuating unconscious biases in research design and outcomes.¹⁰⁵ This suggests that, ideally, teams should be diverse in their make-up and skillsets.

Other studies have recommended diversity and bias education for AI developers and embedding ethicists within AI development teams from the design phase onwards.¹¹¹

Bias against user groups

Bias may occur in ways that affect the users of AI tools (healthcare workers) rather than patients directly. For example, speech recognition algorithms have been shown to be biased on gender¹¹² and race,¹¹³ leading to reduced accuracy. In a healthcare setting, this increased the risk of error and hence has the potential to cause disadvantage or harm for patients based not on their own protected characteristics, but those of their clinician.

Reinforcement of outmoded practice

This bias can occur due to shifts in the landscape of evidence-based best practice. It is specific to recommendation engines (data filtering tools that suggest treatment strategies or care choices). Leveraging retrospective data for AI model training is attractive for practical and economic reasons, but treatment pathways and decisions which were made historically may no longer represent best practice. In this case, models may learn to recommend the out-of-date practice, and even up-to-date algorithms will not adapt to future changes in best practice. Model retraining and continuous evaluation against up-to-date standards is the only effective mitigation for this type of risk, but could be hampered by the lack of available, relevant data in the wake of a change in practice.¹¹⁴

The NHS AI Lab and the Health Foundation are currently supporting several projects to address these issues.¹¹⁵ The projects aim to identify and explore opportunities for using AI to mitigate or overcome existing healthcare inequalities, as well as address the data and model bias issues discussed in **Box 3**. Specifically, the STANdards for Data INclusivity and Generalisability (STANDING together) project aims to address these challenges in part by developing standards for data sets that AI systems use to ensure these systems are diverse, inclusive and work across all demographic groups.⁴³

Interviewees for this research noted that, although bias in AI models is a significant concern, there is no universally accepted method for evaluating healthcare AI products for bias. Established protocols for investigating and mitigating bias during product evaluation and for reporting the potential for bias during deployment may assist with this.

A recent review¹¹⁶ highlights the availability of AI-specific extensions to clinical trial reporting guidelines. These AI-specific guidelines (some listed in **Box 6**) are essential for the systematic assessment of the risk of bias in AI algorithms. Although they are not a measure of bias, they can estimate the risk of undetected bias, given a model and study design. Studies that score high against these guidelines could be considered the gold standard for evidence in healthcare AI research and evaluation.¹¹⁷

Looking beyond bias in AI technologies, human cognitive biases relating to clinical reasoning and decision making are also a challenge, as discussed in detail in **section 5.3**.

An illustration of a person with dark curly hair, wearing a white lab coat over a light blue shirt and dark blue trousers, walking from left to right. The person is positioned to the left of a large yellow rounded rectangle that contains the 'Key confidence insights' section.

Key confidence insights

- » Personal experiences and attitudes to innovation and AI technologies can significantly impact a clinician's confidence in using AI technologies.
- » Clinicians are more likely to trust AI-derived information that does not relate to their area of expertise. This suggests that less experienced clinicians are susceptible to inappropriately high levels of confidence in AI-derived information developed to support their decision making and may require education to critically appraise these technologies.
- » Conversely, experts (specialists in their area of care) tend to be more sceptical and question AI-derived information despite AI technologies having the potential to enhance their decision making performance in some situations.
- » The levels of clinical risk and human involvement associated with specific AI-assisted CRDM scenarios can influence confidence assessment for these technologies.

Key confidence insights (cont.)

- » Clinicians need to be equipped to assess appropriate levels of confidence for each piece of AI-derived information. This assessment should be conducted per patient and with the necessary understanding of the strengths and limitations of the AI technologies involved.
- » Assessing appropriate levels of confidence in AI-derived information requires an understanding of the information that has been considered by the algorithm, and ideally what relative weighting it has been given. However, the technical features of 'black-box' AI technologies can limit such understanding. Further, AI-derived information for CRDM can be qualitatively different to conventional clinical information, both in complexity and the degree of synthesis in the predictions.
- » More research is needed to investigate how certain AI features influence confidence. For example, research is needed to investigate how uncertainty quantification can impact clinical decision making.
- » Explainable AI approaches (XAI) do not currently offer a panacea for assessing confidence in individual AI predictions, and may provide false reassurance to clinicians during CRDM.
- » The potential for bias in AI technologies is a major cause of concern and can affect confidence in these technologies. Established protocols for investigating and mitigating bias during product evaluation and for reporting the potential for bias during deployment may be needed.



5.3 Cognitive biases and appropriate confidence in AI-assisted CRDM

Interactions between humans and systems can be complex as they involve aspects of human psychology.

These complexities are key to understanding confidence in AI-assisted CRDM, particularly in the context of cognitive biases that can occur when humans use shortcuts to make rapid value judgements. These mental shortcuts are necessary to assimilate and evaluate complex information quickly, as described in **5.1.1**.

Interviewees for this research cautioned that, for a new type of information like an AI prediction, mental shortcuts may be inaccurate, and lead to biased value judgments and inappropriate levels of confidence in the information. They stressed

that clinicians will need to identify and mitigate these biases during AI-assisted CRDM, as failure to do so may lead to unnecessary clinical risk. This will depend on a fundamental understanding of cognitive biases and how they apply to CRDM (both with and without AI), which is perceived as a gap in current clinical training and education.

This section describes some of the most common cognitive biases clinicians are susceptible to when using AI-derived information for CRDM. It then discusses how clinical settings, choices in workflow integration and how AI presents information can influence how these biases can adversely affect AI-assisted CRDM.

5.3.1 Cognitive biases affecting AI-assisted CRDM

Several cognitive biases are relevant to AI-assisted CRDM, although this list is not exhaustive:

- » **Automation bias** – the tendency to accept the AI recommendation uncritically, potentially due to time pressure, or under-confidence in the clinical task (for example, in non-specialists). A clinician may tend not to critically appraise AI-derived information in the context of the other available evidence, or without consideration of patient-specific factors affecting the model suitability for this case.
- » **Aversion bias** – the tendency to be sceptical of AI, despite strong evidence that its performance at evaluation was good. An experienced clinician may prefer to rely on tried and tested methods, leading to a tendency to dismiss AI predictions as unnecessary for CRDM, regardless of whether these align or misalign with their clinical judgement.
- » **Alert fatigue** – ignoring alerts provided by an AI system due to history or perception of too many incorrect cases (for example, false positives). This can often be the result of an over-conservatively calibrated AI algorithm, or the users' aversion bias, and occurs often in high volume decision making settings (for example, A&E). Ultimately, the pressure on the downstream resources can overwhelm staff and reduce safety for other patients, making alert-fatigue responses to the AI technology more likely.
- » **Confirmation bias** – accepting AI-derived information uncritically when it agrees with the clinician's intuition, potentially ignoring the evidence suggesting that relying on AI for the specific case should be low. If the AI is operating outside its locus of confidence, this can provide inappropriate reassurance about a decision.
- » **Rejection bias** – rejecting AI recommendations without due consideration when they contradict clinical intuition, and potentially missing the opportunity to critically evaluate whether the AI could have detected something the clinician has not considered.

As noted, interviewees for this research stressed the importance of enabling clinicians to recognise their inherent biases, and to understand how these affect their use of AI-derived information in CRDM, as a key focus of related training and education.

The optimal outcome would be for clinicians to maintain a balanced attitude between enthusiasm for, and healthy scepticism of, AI technologies. Interviewees concluded that the 'critical eye' of an experienced clinician is a vital tool in maintaining patient safety when considering AI-derived information, but it must not become so critical that the value of digital transformation through AI is lost.

5.3.2 Clinical settings and cognitive biases

Feedback from the interviews conducted for this research and analysis of related literature suggest that the impact of the biases described in [section 5.3.1](#) will depend on the clinical setting and use case.

The nature and extent of the task that AI is assigned within a decision making process affect the level of clinical risk associated with that decision (as discussed also in [section 5.2.2](#)). This perceived level of risk can affect a clinician's predisposition towards cognitive biases.¹¹⁸

As described by interviewees for this research, most AI technologies implemented in their settings are limited to decision-support tasks that have low direct clinical risks and maintain human involvement. For example, tools that 'flag' likely serious cases for prioritised radiology reporting still maintain that all images will be read by a human expert.

Nevertheless, interviewees cautioned that the risk of AI adversely influencing human decision making should not be ignored, even in inherently low-risk scenarios. It is possible that human operators come to rely on the lack of an 'AI flag' to indicate low risk, potentially becoming biased towards missing diagnoses when AI flags are absent, particularly when they are under pressure.

Further, in the potential implementation of AI into higher risk CRDM settings, great care will be needed to ensure that the AI is implemented in a way that minimises the risk of cognitive bias for users. This includes users being aware of how aspects of the AI's implementation can impact their decision making.

Requirements for validating AI, as discussed in [section 3.2](#), can mitigate some of this risk. For example, the external validation requirements of the NHS Breast Screening Programme³³ can provide a high level of confidence in AI in inherently high risk, but tightly controlled settings (such as an AI second reader for mammography screening). Existing human reader workflows are designed to mitigate cognitive biases through independent readers operating in isolation, with defined processes for arbitration. AI-enabled workflows following these same approaches will inherently be more robust to cognitive biases. In such an approach, human readers will be less affected by cognitive biases if they are operating or reaching their own conclusions independently of the AI.¹¹⁹

In less formal or more high-pressure situations such as A&E triage, there tends to be stronger polarisation in inherent clinician attitudes to AI (see **section 5.2.1**),¹²⁰ with a higher likelihood that clinicians are strongly in favour or strongly averse to its use. This may be due in part to the rapid nature of the CRDM process in these situations, and the limited opportunity to extensively critique AI-derived information.

Furthermore, as conventional CRDM relies heavily on cognitive shortcuts to assimilate complex information quickly, the introduction of AI into this type of CRDM process is more susceptible to the cognitive biases described earlier.

Interviewees for this research supported this view, perceiving that it would be easy to have inappropriately high confidence in AI technologies that provide an immediate output. However, some workers may have inappropriately low confidence in AI predictions, if they feel they don't understand the technology or aren't convinced by the robustness or generalisability of the evaluation data. In general, interviewees felt that changing these inherent attitudes will be challenging, except in cases of underperformance and clinical incidents.

5.3.3 AI workflow integration and cognitive biases

Feedback from the interviews conducted for this research and analysis of related literature suggest that the design of the workflow and timing of introduction of AI-derived information are key to mitigating cognitive biases.

Automation bias is more likely when the AI-derived information is presented early in the CRDM pathway, as users may cease searching for potentially contradictory evidence if the prediction appears superficially reasonable.¹¹⁴

Without careful workflow design, there is a notable risk that once an AI technology is implemented, evaluated and in successful routine use, users become increasingly uncritical of the AI outputs, resulting in a higher potential for automation bias and clinical error.^{68,89} This is a known phenomenon with any technology embedded in practice, particularly if failure cases are rare, or unreported. As discussed in **section 5.2.1**, less experienced clinicians (for the task in question) will generally be more reliant on AI predictions, making them more susceptible to automation bias than experts, who are more likely to retain a degree of scepticism and be willing to critique the algorithm.¹²¹

One approach to mitigate automation bias, suggested by interviewees for this research, involves delaying the availability of AI-derived information until an initial human opinion has been formed (and recorded). The user would then be required to record whether they have amended their decision or not, in the light of the AI prediction. This has been described in the literature as the 'integrative-AI' approach to clinical decision making.¹²²

In cases where the AI contradicts the initial decision, the clinician is immediately aware of the conflict, whereas with a conventional AI-first approach their own opinion may never have been fully formed. This approach requires a clinician to justify their response to the AI prediction considering their original decision, mitigating the risks of automation and aversion biases.

While suggesting this approach, interviewees recognised that the recording of the initial and final decisions should be as automated and seamless as possible. It should not be overly burdensome to the clinician, who should be focused on the clinical decision, not the record keeping.¹²³ The degree of detail required would be strongly dependent on the clinical scenario.

Research suggests that alert fatigue¹⁰³ can be mitigated by careful calibration of algorithms,¹²⁴ and should be a consideration in the ongoing monitoring of AI-assisted CRDM. The way in which alerts are presented and acknowledged is a key factor in user-interface design for AI systems and can mitigate alert fatigue if done carefully. Designing CRDM systems with prompts and incentives that aim to overcome alert fatigue may be beneficial in these situations.¹²⁵

Confirmation bias becomes a clinical risk when an erroneous human opinion is confirmed by AI, allowing one error to compound another. This has been shown to be more likely when AI-derived information is presented alongside conventional information, allowing the clinician to confirm aspects of the situation and ignore other contradictory factors, which may be critical.¹²² Hence, confirmation bias can be mitigated by presenting the AI prediction later in the pathway, to ensure a thorough human assessment has been performed beforehand, lending further support to the case for integrative-AI workflows.¹²²

Rejection bias is more difficult to mitigate through workflow design, as presenting AI information late in the pathway can increase the likelihood of rejecting it inappropriately.¹²⁶

In order to make AI-assisted CRDM robust against cognitive biases and especially in inherently higher-risk scenarios, interviewees for this research suggested using a decision record system to record and retain the initial clinician assessment, AI prediction and clinician's final decision. This approach is already used in specialties such as radiology, where an initial report is often made and verified by an on-call radiologist in a time-critical scenario, and then a specialist consultant may review and add an addendum, with the initial report retained for reference.

Interviewees noted that a similar approach with AI-assisted CRDM would provide transparency and encourage careful critical assessment of AI-derived information. The full, recorded, decision process would be available for further review, arbitration by an external specialist and, ultimately, any internal investigations or legal process that may occur in case of a clinically significant error. This approach also affords the possibility of continuous feedback, learning and improvement, both for internal teams and AI developers. When clinical incidents do occur and

cause potential or actual patient harm, it has been suggested that cognitive bias assessment should be integrated into the root-cause analysis and reporting for incidents involving AI-assisted CRDM.⁹

The need to record whether a human decision has remained unchanged in the light of contradictory AI information should encourage clinicians to carefully consider their reasons for acceptance or rejection of the AI output and should help to mitigate the risk of the cognitive biases described above.

Finally, this approach would assist in communicating with patients as to how AI has been involved in their care and how the process leading to a clinical decision has happened.

Key confidence insights

- » There are five key cognitive biases that relate to AI-assisted CRDM: Automation bias, aversion bias, alert fatigue, confirmation bias and rejection bias.
- » These biases influence how clinicians assess AI-derived information and the level of confidence that they assign to such information.
- » Enabling clinicians to recognise their inherent biases, and understand how these affect their use of AI-derived information in CRDM should be a key focus of related training and education. Failure to do so may lead to unnecessary clinical risk.
- » The nature and extent of the task that AI is assigned within a decision making process affect the level of clinical risk associated with that decision. This perceived level of risk can affect a clinician's predisposition towards or against each of the cognitive biases.
- » Workflow integration, including the timing of presentation of AI-derived information can influence how clinicians use this information during CRDM.
- » Robust systems for performing and recording AI-assisted CRDM could help clinicians mitigate their potential biases towards AI-derived information.



5.4 Interface with patients

5.4.1 Patient confidence in AI-assisted clinical decisions

Patient attitudes, knowledge levels and responses to the use of AI technologies in their care can vary widely, from apprehension to enthusiasm.¹²⁷

Interviewees for this research perceived that most patients have little to no understanding of the implication of AI use in healthcare, although it should not be assumed that they are not interested and will not have opinions about it. Interested patients should be able to know how the technology they rely on for their health has been developed and what data it relies on. They should also be able to get an explanation from their clinician of how a particular decision has been made, even if the AI itself operates as a 'black-box'. This suggests that different patients will require explanations at various levels of detail.

Interviewees perceived that clinicians should be confident in being able to meet these expectations, including explaining to patients the scope of AI use in their care and the 'checks and balances in place'. They also cautioned that if AI is introduced into CRDM in an opaque way, or if clinicians are not confident in explaining its scope, limitations, benefits and risks, there is a real risk of undermining patient involvement in shared clinical decision making.

5.4.2 The changing relationship between clinicians and patients

The patient-clinician relationship is at the heart of CRDM. It is a key tenet of modern medicine that the patient should be involved in any decision about their care.¹²⁸

As such, the involvement of AI in CRDM can present a 'third-wheel' effect,¹²⁹ in the sense that it is as important for the patient to have confidence in the AI as it is for the clinician, and that the AI may interrupt the patient-clinician relationship. For this reason, it is important that clinicians can take on the role of communicator and educator to their patients, and explain the role and limitations of AI being used in their care.

Interviewees for this research noted that the adoption of AI has the potential to reshape the relationships between clinicians and their patients by introducing further transparency and opportunity for collaboration in shared decision making.

Patients will increasingly have enhanced access to medical knowledge and subsequent decision making for their health, and will require support with assessing probabilities and risks.¹³⁰ Healthcare workers in certain specialties will need to move from being an 'oracle' of clinical information to a health 'counsellor', enabling high-quality, data-driven, shared clinical decision making.⁹

These new dynamics will dictate the required skills for managing the clinician-patient relationship. Clinicians will need to manage the interaction between patients and increasingly complex AI systems, including knowing the limits of

AI and communicating this to patients.¹⁰ This will require an ability to guide the patient through uncertainty around potentially complex diagnostic decisions, and empower patients to take joint responsibility for their healthcare where appropriate.⁹

As healthcare workers interact with ever smarter machines, the demand for soft skills will rise. Social and emotional skills are already becoming more important as technologies take over more physical, repetitive and basic cognitive tasks.⁹

Interviewees concluded, in agreement with previous research, that focusing on the core human skills that AI and computers cannot achieve, such as collaboration, reflection, compassion and empathy will be essential.^{5,130}

Certain specialities such as oncology have great experience of using these human skills to support CRDM, for example, due to the complexity and uncertainty of cancer care. Learning to translate these skills, through education, to other areas where AI will impact CRDM will be beneficial.

Key confidence insights

- » The adoption of AI has the potential to reshape the relationships between clinicians and their patients.
- » Clinicians will need to manage the interaction between patients and increasingly complex AI systems, including knowing the role and limits of AI and communicating this to patients.
- » As healthcare workers and patients interact with ever smarter machines, the demand for soft skills will rise.





Conclusion: Developing Healthcare Workers' Confidence in AI

The main recommendation of this report is to **develop and deploy educational pathways and materials for healthcare professionals at all career points and in all roles, to equip the workforce to confidently evaluate, adopt and use AI**. During clinical decision making, this would enable clinicians to determine appropriate confidence in AI predictions and balance these with other sources of clinical information.

The factors influencing confidence in AI, as detailed in this report, can help to determine the educational requirements to develop such confidence among healthcare workers. The **second report** from this research will outline suggested pathways for related education and training.

Interviewees for this research identified broader efforts that primarily aim to improve patient safety and service delivery, but could also contribute to developing confidence in AI within the healthcare workforce.

Figure 3 shows these efforts mapped across this report's conceptual framework:

- » confidence influenced by the **governance** of AI technologies, with factors relating to regulatory oversight, validation and liability.
- » confidence influenced by the **implementation** of AI technologies at local settings, with factors relating to strategy, culture, IT and IG, local validation, and workflow integration.
- » assessment of appropriate levels of confidence in AI-derived information during **clinical** decision-making, with factors relating to clinician attitudes and cognitive biases, the clinical context, and the AI's features (including explainability).

Many of the identified efforts are already underway, being led by Health Education England, the NHS Transformation Directorate, Integrated Care Systems and trusts, regulators and moderators, legal professionals, academics, and industry innovators.

A forthcoming project will involve engagement with these organisations and relevant groups and sharing of updates on progress being made on these efforts.

Figure 3: Efforts that can contribute towards confidence in AI

Governance

- » Development of professional guidelines on creating, implementing, and using AI for all clinical staff groups
- » Further development of regulatory frameworks for AI performance, quality, and risk management
- » Finalisation of formal requirements for evidence and validation of AI technologies
- » Development of AI specific pathways for prospective clinical studies of new technologies
- » Further development of guidance on liability for AI (including autonomous AI)
- » Establishment of flexible and dynamic processes for developing clinical guidelines on AI-assisted clinical tasks and technologies
- » Development of clear oversight and governance pathways for AI, including AI not classified as a medical device
- » Development of standards for developing AI for health and care settings (including co-creation with users, model transparency and mitigation of model bias)



Implementation

- » Further development of advice, guidelines, and prototypes for information technology (IT) and governance (IG) supporting adoption of AI technologies
- » Development of strategies and assignment of resources to encourage organisational cultures that support innovation, co-creation, and robust appraisal of AI technologies
- » Encouragement of collaboration and sharing of knowledge across NHS sites that are adopting AI technologies
- » Development and resourcing of multi-disciplinary teams across clinical, technical, and administrative roles to enable implementation, local validation, audit and maintenance of AI technologies
- » Establishment of pathways for ongoing monitoring, performance feedback and safety event reporting involving AI technologies



Clinical Use

- » Development of internal systems to record AI-assisted CRDM, including how AI has influenced or changed the decision
- » Further research on explainable AI and its safe use in clinical reasoning and decision making (CRDM)
- » Further research to understand and optimise the presentation of AI-derived information for CRDM
- » Further research to understand how certain AI model features influence confidence
- » Development of confidence in AI technologies across patients and communities via engagement and education activities
- » Support for clinicians to determine appropriate confidence in AI-derived information and balance it with conventional clinical information for CRDM

Appendix A: List of interviewees

Following is a list of individuals interviewed for this research who have agreed to be acknowledged. The report outlines a synthesis of a range of insights and opinions, and as such, responsibility for the content of the report rests with the authors.

Dr Lia Ali
Akrivia Health
Clinical advisor

Dr Nick Barlow
University Hospitals Birmingham NHS Foundation Trust
Director of Applied Digital Healthcare

Nick Capewell
University Hospitals Birmingham NHS Foundation Trust
Lead Ophthalmic Research Technician (Imaging)

Francesca Evans
Aidence
UK Account Manager

Benjamin Fell
Akrivia Health

Dr Claire Fernandez
Brainomix
Clinical Research Programme Manager

Moritz Flockenhaus
CQC
Policy Manager

Tom Gallagher
Dr Doctor
Data Product Manager

Dr Susanne Gaube
University Hospital Regensburg and Ludwig-Maximilians -Universität München
Postdoctoral Research Associate

Dr Saira Ghafur
Imperial College London
Lead for Digital Health, Institute of Global Health Innovation

Maddy Griffiths
ICO
Senior Policy Officer - Innovation Hub

Nik Haliasos
Essex Neurosciences Centre - Redbridge Barking & Havering University Hospitals NHS Trust
Consultant Neurosurgeon

Simon Harris
Kheiron
Senior NHS Project Manager

Dr George Harston
Brainomix
Chief Medical and Innovation Officer

Associate Prof Iain Hennessey
Alder Hey Children's NHS Foundation Trust Hospital
Clinical Director, Alder Hey Innovation Centre

Phillippa Hentsch
University Hospitals Birmingham NHS Foundation Trust
Strategy Lead - Digital Transformation

Dr Caroline Jones
Swansea University
Associate Professor, Law

David King
Aidence
UK Projects and Delivery Manager

Dr Olga Kostopoulou
Imperial College London
Reader in Medical Decision Making

Jeanette Kusel
NICE
Scientific Advice Director

Dr Xiaoxuan Liu
University Hospitals Birmingham NHS Foundation Trust
Clinical Research Fellow

Dr Trystan Macdonald
University Hospitals Birmingham NHS Foundation Trust
Clinical Research Fellow, Specialty Trainee in Ophthalmology

Dr Robert MacLaren
Hammersmith & Fulham GP Partnership
Managing Partner and Caldicott Guardian

Dr Dan Mullarkey
Skin Analytics
Medical Director

Dr Kiruba Nagaratnam
The Royal Berkshire Hospital NHSFT
Clinical Lead for Stroke Medicine

Dr Jonathan Nash
Portsmouth Hospitals NHS Trust
Consultant Breast Radiologist

Dr Luke Nicholson
Moorfields Eye Hospital
Consultant Ophthalmic Surgeon

Damian O'Boyle
Healthy.io
Director of Operations

Johan Ordish
MHRA
Head of Software and AI, Innovative Devices Division

Dr Gurprit Pannu
Sussex Partnership NHS Foundation Trust
Consultant Psychiatrist - Chief Digital Information Officer

Alister Pearson
ICO
Senior Policy Officer - Technology

Dr Russell Pearson
MHRA
NHS X AI Liaison Manager, Innovative Devices Division

Ahmed Razek
ICO
AI specialist and Principal Technology Adviser

Dr Philip Scott
University of Portsmouth
Reader in Health Informatics

Dr Nisha Sharma
Leeds Teaching Hospital NHS Trust
Lead Clinician Radiology

Haris Shuaib
Guy's & St Thomas' NHS Foundation Trust
Consultant Clinical Scientist, Head of Clinical Scientific Computing

Prof Sir David Spiegelhalter FRS OBE
University of Cambridge
Chair, Winton Centre for Risk and Evidence Communication

Harini Suresh
MIT
PhD candidate in Electrical Engineering and Computer Science

Mark Swindells
GMC
Assistant Director - Standards and Ethics

Dr Jay Verma
Shakespeare Health Centre
GP Principal

References

- Joshi I, Morley J. Artificial Intelligence: How to get it right. Putting policy into practice for safe data-driven innovation in health and care. 2019;1-55. <https://www.nhsx.nhs.uk/ai-lab/explore-all-resources/understand-ai/artificial-intelligence-how-get-it-right/%0A> Accessed February 28, 2022.
- Hardie T, Horton T, Willis M, Warburton W. Switched on. How Do We Get the Best out of Automation and AI in Health Care? 2021. doi:10.37829/HF-2021-103
- AI Roadmap report and interactive dashboard - Health Education England. <https://www.hee.nhs.uk/our-work/dart-ed/ai-roadmap>. Accessed February 28, 2022.
- Spiegelhalter D. Should We Trust Algorithms? *Harvard Data Sci Rev*. January 2020;1-12. doi:10.1162/99608f92.cb91a35a
- Topol E. The Topol Review: Preparing the Healthcare Workforce to Deliver the Digital Future. 2019. <https://topol.hee.nhs.uk/the-topol-review/>. Accessed February 28, 2022.
- NHS. NHS Long Term Plan: Digital transformation. NHS England. <https://www.longtermplan.nhs.uk/areas-of-work/digital-transformation/>. Published 2019. Accessed February 28, 2022.
- National AI Strategy - GOV.UK. <https://www.gov.uk/government/publications/national-ai-strategy>. Accessed February 28, 2022.
- The National Strategy for AI in Health and Social Care - NHS AI Lab programmes - NHS Transformation Directorate. <https://www.nhsx.nhs.uk/ai-lab/ai-lab-programmes/the-national-strategy-for-ai-in-health-and-social-care/>. Accessed February 28, 2022.
- Sinha S, Al Huraimel K. Transforming Healthcare with AI. In: *Reimagining Businesses with AI*; 2020:33-54. doi:10.1002/9781119709183.ch3
- Liu X, Keane PA, Denniston AK. Time to regenerate: the doctor in the age of artificial intelligence. *J R Soc Med*. 2018;111(4):113-116. doi:10.1177/0141076818762648
- How to build trust with Trusts on artificial intelligence - Med-Tech Innovation. https://www.med-technews.com/medtech-insights/ai-in-healthcare-insights/how-to-build-trust-with-trusts-on-artificial-intelligence_1/. Accessed February 28, 2022.
- Leslie D. Understanding artificial intelligence ethics and safety. 2019. doi:10.5281/zenodo.3240529
- Parikh RB, Teeple S, Navathe AS. Addressing Bias in Artificial Intelligence in Health Care. *JAMA - J Am Med Assoc*. 2019;322(24):2377-2378. doi:10.1001/jama.2019.18058
- Leslie D, Mazumder A, Peppin A, Wolters MK, Hagerty A. Does "AI" stand for augmenting inequality in the era of covid-19 healthcare? *BMJ*. 2021;372. doi:10.1136/bmj.n304
- UK to pilot world-leading approach to improve ethical adoption of AI in healthcare. GOV.UK. <https://www.gov.uk/government/news/uk-to-pilot-world-leading-approach-to-improve-ethical-adoption-of-ai-in-healthcare>. Accessed March 8, 2022.
- The multi-agency advice service (MAAS) - Regulating the AI ecosystem - NHS Transformation Directorate. <https://www.nhsx.nhs.uk/ai-lab/ai-lab-programmes/regulating-the-ai-ecosystem/the-multi-agency-advice-service-maas/>. Accessed March 7, 2022.
- Digital Technology Assessment Criteria (DTAC) - Key tools and information - NHS Transformation Directorate. <https://www.nhsx.nhs.uk/key-tools-and-info/digital-technology-assessment-criteria-dtac/>. Accessed March 7, 2022.
- MHRA. Software and AI as a Medical Device Change Programme. <https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme/> Published 2021. Accessed March 7, 2022.
- HM Government. Designated standards: medical devices - GOV.UK. <https://www.gov.uk/government/publications/designated-standards-medical-devices>. Accessed March 7, 2022.
- NHS digital, data and technology standards - NHS Digital. <https://digital.nhs.uk/about-nhs-digital/our-work/nhs-digital-data-and-technology-standards>. Accessed March 7, 2022.
- IMDRF. International Medical Device Regulators Forum - Software as a Medical Device (SaMD): Key Definitions. 2013. <https://www.imdrf.org/documents/software-medical-device-samd-key-definitions>. Accessed March 7, 2022.
- CQC. The five key questions we ask - Care Quality Commission. Care Quality Commission. <https://www.cqc.org.uk/what-we-do/how-we-do-our-job/five-key-questions-we-ask>. Published 2016. Accessed March 7, 2022.
- CQC. What we do - Care Quality Commission. <https://www.cqc.org.uk/what-we-do>. Published 2019. Accessed March 7, 2022.
- Richardson JP, Smith C, Curtis S, et al. Patient apprehensions about the use of artificial intelligence in healthcare. *npj Digit Med*. 2021;4(1). doi:10.1038/s41746-021-00509-1
- Wall E, Stasko J, Endert A. Toward a Design Space for Mitigating Cognitive Bias in Vis. 2019 IEEE Vis Conf VIS 2019. 2019:111-115. doi:10.1109/MSUAL.2019.8933611
- Anwar R. Good medical practice. *BMJ*. 2003;327(7425):1213. doi:10.1136/bmj.327.7425.1213
- Smith H. Clinical AI: opacity, accountability, responsibility and liability. *AI Soc*. 2021;36(2):535-545. doi:10.1007/S00146-020-01019-6/FIGURES/1
- Hwang EJ, Park S, Jin KN, et al. Development and Validation of a Deep Learning-Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. *JAMA Netw open*. 2019;2(3):e191095. doi:10.1001/jamanetworkopen.2019.1095
- Beede E, Baylor E, Hersch F, et al. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In: *Conference on Human Factors in Computing Systems - Proceedings*. 2020. doi:10.1145/3313831.3376718
- HM Government. The medical devices regulations 2002. 2002;(618):1-40. <https://www.legislation.gov.uk/ukSI/2002/618/contents/made>. Accessed March 7, 2022.
- Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies in medical imaging. *BMJ*. 2020;368. doi:10.1136/bmj.m689
- Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Heal*. 2019;1(6):e271-e297. doi:10.1016/S2589-7500(19)30123-2

33. Interim guidance on incorporating artificial intelligence into the NHS Breast Screening Programme. Gov.uk. <https://www.gov.uk/government/publications/artificial-intelligence-in-the-nhs-breast-screening-programme/interim-guidance-on-incorporating-artificial-intelligence-into-the-nhs-breast-screening-programme>. Published 2021. Accessed March 7, 2022.
34. Gille F, Jobin A, Ienca M. What we talk about when we talk about trust: Theory of trust for AI in healthcare. *Intell Med*. 2020;1-2:100001. doi:10.1016/j.ibmed.2020.100001
35. NICE. Evidence standards framework for digital health technologies. 2019. <https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies>. Accessed March 7, 2022.
36. NICE. NICE META Tool. <https://meta.nice.org.uk/>. Published 2021. Accessed March 7, 2022.
37. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(7):e048008. doi:10.1136/bmjopen-2020-048008
38. Sounderajah V, Ashrafian H, Golub RM, et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: The STARD-AI protocol. *BMJ Open*. 2021;11(6):e047709. doi:10.1136/bmjopen-2020-047709
39. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI Extension. *BMJ*. 2020;370. doi:10.1136/bmj.m3210
40. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364-1374. doi:10.1038/s41591-020-1034-x
41. Vasey B, Clifton DA, Collins GS, et al. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med*. 2021;27(2):186-187. doi:10.1038/s41591-021-01229-5
42. Sounderajah V, Ashrafian H, Rose S, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nat Med*. 2021;27(10):1663-1665. doi:10.1038/s41591-021-01517-0
43. STANDING Together Working Group. STANDING together. 2021. <https://www.datadiversity.org/>. Accessed March 8, 2022.
44. González-Gonzalo C, Thee EF, Klaver CCW, et al. Trustworthy AI: Closing the gap between development and integration of AI systems in ophthalmic practice. *Prog Retin Eye Res*. December 2021;101034. doi:10.1016/j.preteyeres.2021.101034
45. A buyer's guide to AI in health care - NHS Transformation Directorate. <https://www.nhs.uk/ai-lab/explore-all-resources/adopt-ai/a-buyers-guide-to-ai-in-health-and-care/>. Accessed March 8, 2022.
46. CDDO. Algorithmic Transparency Standard. GOV.UK. <https://www.gov.uk/government/publications/algorithmic-transparency-data-standard>. Published 2021. Accessed March 7, 2022.
47. Google Cloud Model Cards. <https://modelcards.withgoogle.com/about>. Accessed March 7, 2022.
48. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. *npj Digit Med*. 2020;3(1):1-4. doi:10.1038/s41746-020-0253-3
49. Leslie D. Explaining Decisions Made with AI. SSRN Electron J. 2022. doi:10.2139/ssrn.4033308
50. A guide to good practice for digital and data-driven health technologies - GOV.UK. <https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology>. Accessed March 7, 2022.
51. What Good Looks Like framework - What Good Looks Like - NHS Transformation Directorate. <https://www.nhs.uk/digitise-connect-transform/what-good-looks-like/what-good-looks-like-publication/>. Accessed March 7, 2022.
52. A guide to using artificial intelligence in the public sector - GOV.UK. <https://www.gov.uk/government/publications/a-guide-to-using-artificial-intelligence-in-the-public-sector>. Accessed March 7, 2022.
53. Good Machine Learning Practice for Medical Device Development: Guiding Principles - GOV.UK. <https://www.gov.uk/government/publications/good-machine-learning-practice-for-medical-device-development-guiding-principles>. Accessed March 7, 2022.
54. Medical Technologies Evaluation Programme - NICE guidance. <https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-medical-technologies-evaluation-programme>. Accessed March 7, 2022.
55. Diagnostics Assessment Programme - NICE guidance - Our programmes. <https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-diagnostics-guidance>. Accessed March 7, 2022.
56. HeartFlow FFRCT for estimating fractional flow reserve from coronary CT angiography - Guidance - NICE. <https://www.nice.org.uk/guidance/mtg32>. Accessed March 7, 2022.
57. Zio XT for detecting cardiac arrhythmias - Guidance - NICE. <https://www.nice.org.uk/guidance/mtg52>. Accessed March 7, 2022.
58. An Innovator's Guide to the NHS.; 2020. https://www.boehringer-ingelheim.co.uk/sites/gb/files/documents/innovators_guide.pdf. Accessed March 7, 2022.
59. NICE. Medtech innovation briefings. <https://www.nice.org.uk/about/what-we-do/our-programmes/nice-advice/medtech-innovation-briefings>. Accessed March 7, 2022.
60. NICE. The technologies - Artificial intelligence in mammography. <https://www.nice.org.uk/advice/mib242/chapter/The-technologies>. Accessed March 7, 2022.
61. Principled Artificial Intelligence - Berkman Klein Center. <https://cyber.harvard.edu/publication/2020/principled-ai>. Published 2020. Accessed March 7, 2022.
62. Government Digital Service. Data Ethics Framework - GOV.UK. Government Digital Service. <https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework-2020>. Published 2020. Accessed March 7, 2022.
63. WHO. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance.; 2021. <http://apps.who.int/bookorders>. Accessed March 7, 2022.
64. Hesketh R. Trusted autonomous systems in healthcare A policy landscape review. 2021. doi:10.18742/pub01-062
65. NHS AI Virtual Hub - NHS Transformation Directorate. <https://www.nhs.uk/ai-lab/ai-lab-virtual-hub/>. Accessed March 8, 2022.
66. Dermatology digital playbook - Digital playbooks - NHS Transformation Directorate. <https://www.nhs.uk/key-tools-and-info/digital-playbooks/dermatology-digital-playbook/>. Accessed March 7, 2022.
67. NHS. Interoperability Toolkit - NHS Digital. <https://digital.nhs.uk/services/interoperability-toolkit>. Published 2021. Accessed March 7, 2022.

68. Gaube S, Suresh H, Raue M, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digit Med.* 2021;4(1):1-8. doi:10.1038/s41746-021-00385-9
69. Garcia-Vidal C, Sanjuan G, Puerta-Alcalde P, Moreno-García E, Soriano A. Artificial intelligence to support clinical decision making processes. *EBioMedicine.* 2019;46:27-29. doi:10.1016/j.ebiom.2019.07.019
70. van Baalen S, Boon M, Verhoef P. From clinical decision support to clinical reasoning support systems. *J Eval Clin Pract.* 2021;27(3):520-528. doi:10.1111/jep.13541
71. NICE. NICE guidelines. PSA testing | Diagnosis | Prostate cancer | CKS |. <https://cks.nice.org.uk/topics/prostate-cancer/diagnosis/psa-testing/>. Published 2017. Accessed February 28, 2022.
72. Saraiya M, Kottiri BJ, Leadbetter S, et al. Total and percent free prostate-specific antigen levels among U.S. men, 2001-2002. *Cancer Epidemiol Biomarkers Prev.* 2005;14(9):2178-2182. doi:10.1158/1055-9965.EPI-05-0206
73. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 2019;17(1):1-9. doi:10.1186/s12916-019-1426-2
74. Magrabi F, Ammenwerth E, McNair JB, et al. Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications. *Yearb Med Inform.* 2019;28(1):128-134. doi:10.1055/s-0039-1677903
75. Myers PD, Ng K, Severson K, et al. Identifying unreliable predictions in clinical risk models. *npj Digit Med.* 2020;3(1):1-8. doi:10.1038/s41746-019-0209-7
76. Benda NC, Novak LL, Reale C, Ancker JS. Trust in AI: why we should be designing for APPROPRIATE reliance. *J Am Med Inform Assoc.* 2021;29(1):207-212. doi:10.1093/jamia/ocab238
77. Shen J, Zhang CJP, Jiang B, et al. Artificial intelligence versus clinicians in disease diagnosis: Systematic review. *JMIR Med Informatics.* 2019;7(3):e10010. doi:10.2196/10010
78. Lee MH, Siewiorek DP, Smalagic A. A human-ai collaborative approach for clinical decision making on rehabilitation assessment. *Conf Hum Factors Comput Syst - Proc.* 2021;(Figure 1). doi:10.1145/3411764.3445472
79. Asan O, Bayrak AE, Choudhury A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *J Med Internet Res.* 2020;22(6):e15154. doi:10.2196/15154
80. Petkus H, Hoogewerf J, Wyatt JC. What do senior physicians think about AI and clinical decision support systems: Quantitative and qualitative analysis of data from specialty societies. *Clin Med J R Coll Physicians London.* 2020;20(3):324-328. doi:10.7861/clinmed.2019-0317
81. Westbrook JI, Raban MZ, Walter SR, Douglas H. Task errors by emergency physicians are associated with interruptions, multitasking, fatigue and working memory capacity: A prospective, direct observation study. *BMJ Qual Saf.* 2018;27(8):655-663. doi:10.1136/bmjqs-2017-007333
82. Larasati R, Liddo A De, Motta E. AI Healthcare System Interface: Explanation Design for Non-Expert User Trust. *CEUR Workshop Proc.* 2021;2903.
83. Macrae C. Governing the safety of artificial intelligence in healthcare. *BMJ Qual Saf.* 2019;28(6):495-498. doi:10.1136/bmjqs-2019-009484
84. Blease C, Bernstein MH, Gaab J, et al. Computerization and the future of primary care: A survey of general practitioners in the UK. *PLoS One.* 2018;13(12):e0207418. doi:10.1371/journal.pone.0207418
85. PWC. What doctor? What Dr. 2017;(June):1-50. <http://medicalfuturist.com/>. Accessed February 28, 2022.
86. Mori I. Public views of Machine Learning Findings from public research and engagement. 2017;(April). <http://www.ipsos-mori.com/terms>. Accessed February 28, 2022.
87. Holm S. Handle with care: Assessing performance measures of medical AI for shared clinical decision making. *Bioethics.* 2022;36(2):178-186. doi:10.1111/bioe.12930
88. Bond RR, Mulvenna M, Wang H. Human centered artificial intelligence: Weaving UX into algorithmic decision making. *RoCHI 2019 Int Conf Human-Computer Interact.* 2019:2-9. <https://hai.stanford.edu>. Accessed March 8, 2022.
89. Buçinca Z, Malaya MB, Gajos KZ. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision making. 2021;5(April). doi:10.1145/3449287
90. Cai CJ, Winter S, Steiner D, Wilcox L, Terry M. "Hello AI": Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision making. *Proc ACM Human-Computer Interact.* 2019;3(CSCW). doi:10.1145/3359206
91. Chari S, Seneviratne O, Gruen DM, Foreman MA, Das AK, McGuinness DL. Explanation Ontology: A Model of Explanations for User-Centered AI. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 12507 LNCS. Springer Science and Business Media Deutschland GmbH; 2020:228-243. doi:10.1007/978-3-030-62466-8_15
92. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: *34th International Conference on Machine Learning, ICML 2017*. Vol 3. ; 2017:2130-2143.
93. Zhang Y, Vera Liao Q, Bellamy RKE. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ; 2020:295-305. doi:10.1145/3351095.3372852
94. Cutillo CM, Sharma KR, Foschini L, et al. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. *npj Digit Med.* 2020;3(1):1-5. doi:10.1038/s41746-020-0254-2
95. Watson D. The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence. *Minds Mach.* 2019;29(3):417-440. doi:10.1007/s11023-019-09506-6
96. Winkler JK, Fink C, Toberer F, et al. Association between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatology.* 2019;155(10):1135-1141. doi:10.1001/jamadermatol.2019.1735
97. Tjoa E, Guan C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Trans Neural Networks Learn Syst.* 2021;32(11):4793-4813. doi:10.1109/TNNLS.2020.3027314
98. Jin W, Li X, Hamarneh G. One Map Does Not Fit All: Evaluating Saliency Map Explanation on Multi-Modal Medical Images. July 2021. <https://arxiv.org/abs/2107.05047v1>. Accessed February 28, 2022.
99. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. In: *Advances in Neural Information Processing Systems*. Vol 2018-Decem. Neural information processing systems foundation; 2018:9505-9515. <https://arxiv.org/abs/1810.03292v3>. Accessed February 28, 2022.
100. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Heal.* 2021;3(11):e745-e750. doi:10.1016/s2589-7500(21)00208-9
101. Babic BB, Gerke S, Evgeniou T, Glenn Cohen I. Beware explanations from AI in health care the benefits of explainable artificial intelligence are not what they appear. *Science.* 2021;373(6552):284-286. doi:10.1126/science.abg1834

102. Chen C, Li O, Tao C, Barnett AJ, Su J, Rudin C. This looks like that: Deep learning for interpretable image recognition. In: *Advances in Neural Information Processing Systems*. Vol 32. Neural information processing systems foundation; 2019. <https://arxiv.org/abs/1806.10574v5>. Accessed February 28, 2022.
103. Yu KH, Kohane IS. Framing the challenges of artificial intelligence in medicine. *BMJ Qual Saf*. 2019;28(3):238-241. doi:10.1136/bmjqs-2018-008551
104. Cho MK. Rising to the challenge of bias in health care AI. *Nat Med*. 2021;27(12):2079-2081. doi:10.1038/s41591-021-01577-2
105. Zou J, Schiebinger L. Ensuring that biomedical AI benefits diverse populations. *EBioMedicine*. 2021;67. doi:10.1016/j.ebiom.2021.103358
106. Center for data ethics and innovation. Review into bias in algorithmic decision making Centre for Data Ethics and Innovation. 2020
107. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453. doi:10.1126/science.aax2342
108. Henry Kamulegeya L, Okello M, Mark Bwanika J, et al. Using artificial intelligence on dermatology conditions in Uganda: A case for diversity in training data sets for machine learning. *bioRxiv*. October 2019:826057. doi:10.1101/826057
109. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ*. 2020;369:26. doi:10.1136/bmj.m1328
110. Subbaswamy A, Adams R, Saria S. Evaluating Model Robustness and Stability to Data set Shift. 2020;130. <http://arxiv.org/abs/2010.15100>. Accessed February 28, 2022.
111. McLennan S, Fiske A, Celi LA, et al. An embedded ethics approach for AI development. *Nat Mach Intell*. 2020;2(9):488-490. doi:10.1038/s42256-020-0214-1
112. Tatman R. Gender and Dialect Bias in YouTube's Automatic Captions. In: *EACL 2017 - Ethics in Natural Language Processing, Proceedings of the 1st ACL Workshop*. 2017:53-59. doi:10.18653/v1/w17-1606
113. Koenecke A, Nam A, Lake E, et al. Racial disparities in automated speech recognition. *Proc Natl Acad Sci U S A*. 2020;117(14):7684-7689. doi:10.1073/pnas.1915768117
114. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*. 2019;28(3):231-237. doi:10.1136/bmjqs-2018-008370
115. The AI Ethics Initiative - NHS AI Lab programmes - NHS Transformation Directorate. <https://www.nhs.uk/ai-lab/ai-lab-programmes/ethics/>. Accessed March 8, 2022.
116. Ibrahim H, Liu X, Denniston AK. Reporting guidelines for artificial intelligence in healthcare research. *Clin Exp Ophthalmol*. 2021;49(5):470-476. doi:10.1111/ceo.13943
117. McCradden MD, Joshi S, Anderson JA, Mazwi M, Goldenberg A, Shaul RZ. Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. *J Am Med Informatics Assoc*. 2020;27(12):2024-2027. doi:10.1093/jamia/ocaa085
118. Kliegr T, Bahník Š, Fürnkranz J. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artif Intell*. 2021;295:103458. doi:10.1016/j.artint.2021.103458
119. Hickman SE, Baxter GC, Gilbert FJ. Adoption of artificial intelligence in breast imaging: evaluation, ethical constraints and limitations. *Br J Cancer*. 2021;125(1):15-22. doi:10.1038/s41416-021-01333-w
120. Stewart J, Sprivilis P, Dwivedi G. Artificial intelligence and machine learning in emergency medicine. *EMA - Emerg Med Australas*. 2018;30(6):870-874. doi:10.1111/1742-6723.13145
121. Goddard K, Roudsari A, Wyatt JC. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *J Am Med Informatics Assoc*. 2012;19(1):121-127. doi:10.1136/amiajnl-2011-000089
122. Braun M, Hummel P, Beck S, Dabrock P. Primer on an ethics of AI-based decision support systems in the clinic. *J Med Ethics*. 2021;47(12):E3. doi:10.1136/medethics-2019-105860
123. Dymek C, Kim B, Melton GB, Payne TH, Singh H, Hsiao CJ. Building the evidence-base to reduce electronic health record-related clinician burden. *J Am Med Inform Assoc*. 2021;28(5):1057-1061. doi:10.1093/jamia/ocaa238
124. Co Z, Holmgren AJ, Classen DC, et al. The tradeoffs between safety and alert fatigue: Data from a national evaluation of hospital medication-related clinical decision support. *J Am Med Informatics Assoc*. 2020;27(8):1252-1258. doi:10.1093/jamia/ocaa098
125. Medlock S, Wyatt JC, Patel VL, Shortliffe EH, Abu-Hanna A. Modeling information flows in clinical decision support: Key insights for enhancing system effectiveness. *J Am Med Informatics Assoc*. 2016;23(5):1001-1006. doi:10.1093/jamia/ocv177
126. Burton JW, Stein MK, Jensen TB. A systematic review of algorithm aversion in augmented decision making. *J Behav Decis Mak*. 2020;33(2):220-239. doi:10.1002/bdm.2155
127. Young AT, Amara D, Bhattacharya A, Wei ML. Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. *Lancet Digit Heal*. 2021;3(9):e599-e611. doi:10.1016/S2589-7500(21)00132-1
128. De Silva D. Helping people share decision making | The Health Foundation. The Health Foundation. <https://www.health.org.uk/publications/helping-people-share-decision-making>. Published 2012. Accessed March 7, 2022.
129. Triberti S, Durosini I, Pravettoni G. A "Third Wheel" Effect in Health Decision Making Involving Artificial Entities: A Psychological Perspective. *Front Public Heal*. 2020;8(April):1-9. doi:10.3389/fpubh.2020.00117
130. Building a Smarter Health Care Workforce Using AI. *AHA Cent Heal Innov*. 2019. https://www.aha.org/system/files/media/file/2019/09/Market_Insights_AI_Workforce_2.pdf. Accessed March 7, 2022.

NHS AI Lab / NHS Transformation Directorate

- » NHS AI Lab: www.nhsx.nhs.uk/ai-lab/
- » Explore all AI resources: www.nhsx.nhs.uk/ai-lab/explore-all-resources/
- » Community of practice AI Virtual Hub: www.nhsx.nhs.uk/ai-lab/ai-lab-virtual-hub/
- » Twitter: @NHSTransform www.twitter.com/NHSTransform
- » LinkedIn: www.linkedin.com/company/nhstransform/posts/

Health Education England

- » Health Education England: <http://www.hee.nhs.uk/>
- » HEE Digital Transformation: <https://digital-transformation.hee.nhs.uk/>
- » Twitter: @HEE_DigiReady www.twitter.com/NHSTransform
- » Email: dart-ed@hee.nhs.uk

