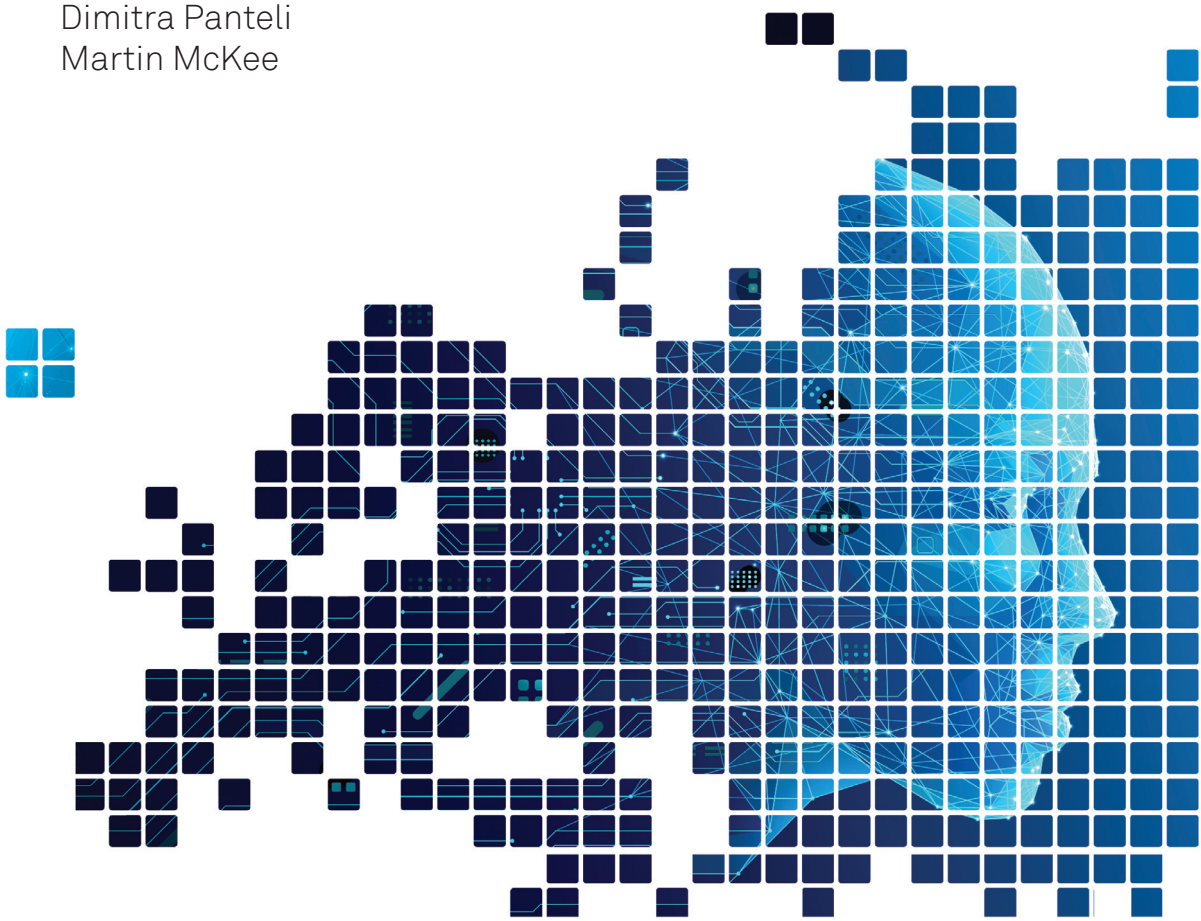
Demystifying artificial intelligence in health

63

Health Policy Series

What health policy-makers need to know

Paula del Rey Puech
Jasjot Saund
Dimitra Panteli
Martin McKee



European Observatory on Health Systems and Policies
a partnership hosted by WHO

Demystifying artificial intelligence in health

What health policy-makers need to know

The European Observatory on Health Systems and Policies supports and promotes evidence-based health policy-making through comprehensive and rigorous analysis of health systems in Europe. It brings together a wide range of policy-makers, academics and practitioners to analyse trends in health reform, drawing on experience from across Europe to illuminate policy issues.

The Observatory is a partnership, hosted by the WHO Regional Office for Europe, which includes other international organizations (the European Commission); national and regional governments (Austria, Belgium, Finland, Ireland, the Netherlands (Kingdom of the), Norway, Slovenia, Spain, Sweden, Switzerland, the United Kingdom and the Veneto Region of Italy with Agenas); other health system organizations (the French National Union of Health Insurance Funds (UNCAM), the Health Foundation); and academia (the London School of Economics and Political Science (LSE) and the London School of Hygiene & Tropical Medicine (LSHTM)).The Observatory has a secretariat in Brussels and it has hubs in London at LSE and LSHTM) and at the Berlin University of Technology.

Demystifying artificial intelligence in health

What health policy-makers need to know

Paula del Rey Puech
London School of Hygiene & Tropical Medicine and Royal Free London NHS Trust

Jasjot Saund
Royal Free London NHS Trust and AI Centre for Value Based Healthcare

Dimitra Panteli
European Observatory on Health Systems and Policies

Martin McKee
European Observatory on Health Systems and Policies

European Observatory on Health Systems and Policies
a partnership hosted by WHO

Keywords:

DIGITAL HEALTH
TELEMEDICINE
INTEGRATION
HEALTH POLICY

Suggested citation. del Rey Puech P, Saund J, Panteli D, McKee M. Demystifying artificial intelligence in health: what health policy-makers need to know. Copenhagen: European Observatory on Health Systems and Policies, WHO Regional Office for Europe; 2026. Licence: CC BY-NC-SA 3.0 IGO.

Cataloguing-in-Publication (CIP) data. CIP data are available at http://apps.who.int/iris.

Sales, rights and licensing. To purchase WHO publications, see https://www.who.int/publications/book-orders. To submit requests for commercial use and queries on rights and licensing, please contact contact@obs.who.int.

ISBN 9789289014755 (electronic version)
ISBN 9789289014786 (print version)

Printed in the United Kingdom

Contents

List of figures, tables and boxes

Figures

Tables

Boxes

Acknowledgements

The main authors of this publication are Paula del Rey Puech (London School of Hygiene & Tropical Medicine and Royal Free London, NHS Trust), Jasjot Saund (Royal Free London NHS Trust and AI Centre for Value Based Healthcare), Dimitra Panteli (European Observatory on Health Systems and Policies) and Martin McKee (European Observatory on Health Systems and Policies).

The authors are grateful to Sanjay Kinra, Hajdi Kosednar, Katherina Ladewig, Jukka Lähesmaa, Rikard Rosenbacke, Liisa-Maria Voipio-Pulkki and Varthani Kirupanandan for undertaking detailed reviews of the final draft and for their very valuable comments. Any remaining errors are the responsibility of the authors. The authors benefited greatly from discussions with Rikard Rosenbacke, who has shared his often profound thinking on the nature of artificial intelligence with us.

The authors are especially grateful to those who contributed to the case studies: Jukka Lähesmaa (Ministry of Social Affairs and Health of Finland) and Tuomas Pöysti (former Chancellor of Justice, Finland); Katharina Ladewig (Centre for Artificial Intelligence in Public Health Research, Robert Koch Institute, Berlin, Germany); Hajdi Kosednar (Health Care Informatics Centre, National Institute of Public Health of Slovenia, Ljubljana, Slovenia) and Vesna Kerstin Petrič (Office for Cooperation with WHO, Ministry of Health of Slovenia, Ljubljana, Slovenia); Tino Martí (Department of Health in the Autonomous Community of Catalonia, Spain) and Oscar Solans (the public health service of the Autonomous Community of Catalonia, Spain (CatSalut)); and to Liisa-Maria Voipio-Pulkki for facilitating links with Finnish colleagues.

Abbreviations

AI	artificial intelligence
AUC-ROC	area under the receiver operating characteristic curve
BLEU	Bilingual Evaluation Understudy
COVID-19	coronavirus disease
CNN	convolutional neural network
CT	computed tomography
EU	European Union
FDP	Federated Data Platform
GAN	generative adversarial network
GDPR	General Data Protection Regulation
GeoAI	geospatial artificial intelligence
GIS	geographic information system
HTA	health technology assessment
ICO	Information Commissioner's Office
IT	Information technology
IVDR	In Vitro Diagnostic Medical Devices Regulation
LLM	large language model
LMM	large multimodal model
MDR	Medical Device Regulation
MLOps	machine learning operations
MRI	magnetic resonance imaging
NHS	National Health Service (United Kingdom)

NICE	National Institute for Health and Care Excellence (United Kingdom)
NLP	natural language processing
PEMAT	Patient Education Materials Assessment Tool
RAG	retrieval-augmented generation
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SDE	secure data environment
UNESCO	United Nations Educational, Scientific and Cultural Organization
WHO	World Health Organization
XAI	explainable artificial intelligence

Foreword

As the outgoing and incoming Chairs of the European Observatory's Steering Committee, we are pleased to introduce this important book on artificial intelligence (AI) in health, which provides timely, practical guidance to help professionals and decisionmakers navigate the opportunities and challenges presented by AI. AI is no longer a distant prospect for European health systems. It is already reshaping how we deliver and govern health.

From our vantage points, we have watched European health systems navigate both promise and uncertainty. AI tools are beginning to support diagnoses, guide clinical pathways, strengthen surveillance and help public health authorities anticipate risks in ways that were unimaginable a decade ago. As in other sectors and society in general, these advances bring profound questions about safety, transparency, fairness and accountability. This book addresses these issues head-on, translating a complex technical field into clear insights for health policy-makers, planners and practitioners. Three themes are especially salient.

First, the European Union's (EU's) regulatory architecture, from the AI Act and the General Data Protection Regulation to the Medical Device Regulation and the European Health Data Space, has been built on a simple truth: technology must serve people, not the other way around. These regulatory frameworks reflect the European values of dignity, democracy and equity. They will only succeed, however, if citizens and professionals can trust that AI tools are safe, reliable and subject to meaningful human oversight. Trust is earned through evidence of real-world performance, transparency in design and vigilance against bias and model drift. This book clarifies what such evidence entails.

Second, AI's transformative potential extends far beyond the clinical setting. The ability to integrate diverse data sources, on the environment, housing, transport, demography and social determinants, opens new possibilities for prevention and prediction. AI can help health systems shift from reactive treatment to proactive, population-level action. It can uncover patterns that shape inequities and sharpen our ability to intervene early. But realizing this potential requires investment in interoperable infrastructure, a skilled workforce and robust governance, so that innovations that begin as isolated pilots can become sustainable components of resilient health systems.

Third, AI's implementation must enhance equity. Bias can seep into AI systems at every stage, from data collection to model training to deployment, and can harm those already facing disadvantages. Ensuring fairness demands co-creation with communities, inclusive design and credible accountability mechanisms.

We share confidence in Europe's ability to steer AI development responsibly. Europe benefits from comprehensive health systems based on the principle of solidarity, strong public institutions, committed researchers, engaged civil societies and a long-standing tradition of dialogue. The European Observatory's role in generating independent analysis, fostering shared learning and supporting evidence-informed policy-making has never been more essential. This book embodies that mission.

Looking ahead, wise implementation of AI can enhance the quality, safety, efficiency and equity of European health systems. It can strengthen preparedness for future shocks and support healthier lives across Europe. But this will require the humility to recognize the limitations of the technology, and the ambition to ensure that technological progress aligns with public values and societal needs.

We therefore commend this book to all those shaping the future of health in Europe. May it support thoughtful decision-making, inspire responsible innovation and encourage the continued cooperation that lies at the heart of the European Observatory's work.

Liisa-Maria Voipio-Pulkki, Ministry of Health and Social Affairs, Finland (ret.)

Stefan Eichwalder, Federal Ministry of Labour, Social Affairs, Health and Consumer Protection, Austria

What is this book about, and who is it for?

This book has been written for those working in, or with, the health sector who will, increasingly, have to engage with the topic of artificial intelligence (AI). We know, from the extensive discussions that we have had before and during our writing of this book, that they will come to this topic from very different starting points. Some may be encountering these concepts for the first time, while others may already have technical or clinical experience with them (or, more often, some of them). As a result, their needs and expectations will vary widely, and this book aims to provide a foundation that is accessible and relevant to a wide variety of readers, helping all readers engage meaningfully with the issues at hand.

This is far from easy. This is a topic that is complex and evolving rapidly. While the chapters that follow will explore the issues in much greater detail, we have written this foreword as a high-level summary of the key messages about the use of AI in health. It aims to provide a balanced overview of the opportunities, risks and considerations that come with integrating AI into a wide range of health sector activities. In the following paragraphs we summarize the key messages that have emerged from the book, although, inevitably, these have been simplified and interested readers should turn to the corresponding sections to understand the issues in more detail.

The first point is that we must understand that AI is not a single technology. Although public attention is currently focused on large language models (LLMs), such as those powering ChatGPT and Gemini, AI encompasses a wide range of techniques, including traditional machine learning, deep learning and computer vision (please see the Executive summary for core concepts in AI and useful definitions). Each of these techniques has different applications and implications for health. For example, traditional machine learning and deep learning have long been explored in areas of health such as medical imaging and risk prediction, but the emergence of generative AI has introduced new possibilities, and also new concerns.

The excitement surrounding AI has led to two dominant narratives: one that sees it as a panacea for all our challenges, and another that warns of catastrophic

consequences. Although there is some debate, many experts see AI as a tool and, like any tool, its impact depends on how it is used. One thing we have come to appreciate, although rarely discussed in the literature, is that the value of AI often depends on the skills and motivations of those using it. Nuclear fusion can generate the energy we need to power our economies, or it can devastate cities when used in a bomb. Also, while some argue that AI now has agency and will increasingly make decisions for and about us, the consensus remains that human decisions shape its outcomes. This means that ethical, informed and context-sensitive use is essential.

Despite its widely acknowledged potential, AI adoption in health is slow and still in its early stages. Significant barriers include the risk to patient safety if AI systems fail and the black box nature of deep learning models. As a result, most of the emerging applications of generative AI are focused on operational tasks, such as transcription and summarization, where the stakes are lower. Clinical decision-support tools using generative AI are also being researched and piloted, but they typically require human oversight to ensure safety and accountability.

While AI's role in health is often framed around clinical care, where its development may be most advanced, its potential reaches well beyond direct clinical applications. In public health, for example, it offers transformative potential by enabling integrated, data-driven insights into population health. Beyond improving efficiency, it can reveal hidden patterns in data to inform targeted interventions, enhance disease forecasting and support behavioural change. By leveraging diverse data sources and offering advanced analytical techniques, AI can help deepen our understanding of health determinants, support segmentation and risk stratification, and strengthen communication and public engagement. However, as with other technological innovations, these applications require interdisciplinary collaboration, careful interpretation and evaluation, and thoughtful consideration of the risks and opportunities posed by different AI models and tools compared to existing approaches.

AI is also making important indirect contributions to health through improvements in logistics, management and operational efficiency. It is being used, or explored, to optimize hospital workflows, streamline administrative processes, such as scheduling and billing, and manage supply chains more effectively. These behind-the-scenes uses can help reduce costs, minimize delays and free-up health care professionals to focus more on patient care, with the goal of supporting better outcomes across the system.

Some caveats are necessary. Data quality is a critical issue. Many predictive models rely on electronic health records to train them and these are often incomplete or episodic. This limits their accuracy and usefulness. Long-term

investment in health care data infrastructure is needed to provide high-quality data for training and updating AI models. Without this, there is a risk of relying on models trained in other settings, which may not be suitable for the target population. Bias is a related concern and can emerge throughout the AI development and implementation life-cycle, such as in the training datasets, which can lead to biased outputs, exacerbating disparities in care. Differences in language, culture and demographics must also be accounted for to ensure AI systems serve diverse populations fairly. This is especially important in public health communication, where inaccurate or misleading content can harm public safety and undermine trust.

The history of AI shows that there is often a lag between theoretical breakthroughs and practical implementation. For instance, neural networks were conceptualized in the last century, but only became viable decades afterwards with advances in computing power and data availability. Similarly, the transformer architecture that underpins many LLMs was introduced in 2017, but commercial large-scale models only emerged recently. This highlights the importance of infrastructure, resources and adoption pathways in realizing AI's potential. It is also important to separate the hype from the reality and to view AI through a historical lens, recognizing the recurring cycles of optimism and disillusionment. Many of the early hopes, for example, in drug discovery, have yet to be realized. While some experts argue that today's innovations mark a fundamental shift, particularly in generative and multimodal AI, whether this cycle will deliver on all of its promises remains to be seen.

This takes us to trust. Overstated claims can lead to misplaced trust or fear, while a clear-eyed understanding of what AI can and cannot do helps guide its safe, effective and ethical integration into the health sector. Trust must be earned. This involves both cognitive trust, based on rational evaluation of reliability, and affective trust, shaped by users' experiences and emotions. Developers must communicate clearly about what AI can and cannot do, and users need education to assess AI outputs critically, particularly in the context of the generative AI hallucination risks. Blind faith or excessive scepticism are both problematic; a balanced approach is required.

Regulatory bodies and technology assessment organizations face the serious challenge of keeping up with fast-moving developments. They must recruit and retain experts in AI, collaborate with relevant institutions and avoid conflicts of interest. Ethical use of AI requires robust governance frameworks, technical safeguards and sustained human oversight. The European Union (EU) AI Act is one example of a risk-based regulatory approach aimed at promoting transparency and mitigating bias. Real-world evaluation, implementation and monitoring

of AI tools are essential to ensure they work safely and effectively in practice. More fundamentally, we must consider how human values can be integrated into increasingly autonomous AI systems.

One of the most pressing questions is how AI should be integrated into clinical workflows. Should it act as a stand-alone diagnostic tool, or support clinicians in decision-making? Most experts agree that human oversight is necessary, especially given the risks for patient safety if AI were to fail. However, humans also make mistakes and errors can arise in machine–human collaboration. These errors can take many forms and must be carefully studied to improve safety and reliability.

The dual-use nature of AI presents a regulatory challenge. Tools developed for beneficial research can be repurposed for harm, such as designing biological weapons. This has prompted calls for international cooperation, technical safeguards and agile governance to prevent misuse and enhance biosecurity.

In clinical research, AI tools used in evidence synthesis may produce plausible but incorrect outputs. Expert oversight is essential to ensure reliability. Moreover, AI tools should be evaluated against current best practice, not outdated or poor-quality care, to assess their effectiveness accurately. Guidelines such as SPIRIT-AI and CONSORT-AI extend existing clinical trial standards to address the unique challenges of AI-based interventions, including dynamic algorithm behaviour and human–AI interaction.

Integrating AI into clinical workflows requires infrastructure that supports real-time data exchange, model interpretability and feedback from health professionals. This ensures that AI supports, rather than disrupts, care delivery. Unlocking AI's potential also depends on education and training systems that prepare workers for the future, including digital literacy and lifelong learning. Safety nets must be in place to support those affected by automation.

Finally, the responsibility gap created by opaque AI systems poses legal and ethical challenges. Traditional notions of liability may not apply, making it difficult for individuals to seek justice when harm occurs. Responsible governance and appropriate regulation are essential to ensure that AI serves human well-being without becoming a tool of unchecked surveillance.

In summary, AI offers exciting possibilities for improving the health of populations and individuals, and its deployment must be grounded in the appropriate technical, ethical and governance foundations. Many experts believe that the hype will eventually settle (some say it already has) and AI will become one tool among many. The challenge is to use it wisely, ensuring that it enhances care, promotes equity and protects patient safety.

Executive summary

AI holds immense promise for improving health outcomes, enhancing system efficiency and empowering both professionals and patients. Advances in computing power, data availability and algorithmic sophistication, particularly in deep learning and generative AI, have enabled machines to perform increasingly complex tasks. To fully realize these benefits, however, the rapid pace of innovation must be matched by regulatory frameworks and real-world evidence to manage emerging risks. Yet, while AI can streamline operations and support decision-making, its adoption in the health sector has been uneven. While enthusiasm for AI is high, many claims about AI's capabilities are overstated, driven by commercial incentives and not supported by long-term clinical evidence. Clearer communication and rigorous validation can help bridge this gap and strengthen patient safety, public trust and system integrity.

AI systems can perpetuate and amplify existing biases, especially when trained on unrepresentative data. This can lead to discriminatory outcomes, particularly for historically underserved populations. Generative AI tools can produce so-called hallucinations, producing plausible but incorrect or misleading information, which is especially dangerous in clinical and public health contexts. Regulatory and technical approaches are essential to mitigate these risks. Overreliance on automated system's outputs, known as automation bias, can further compromise decision-making and accountability.

Health systems often lack the infrastructure, workforce capacity and governance mechanisms needed to deploy AI safely and effectively. Without secure data environments, interoperable systems and continuous monitoring, AI tools may degrade over time due to model and data drift, leading to unreliable or unsafe outputs. The environmental impact of large-scale AI systems, including high energy and water consumption, also raises sustainability concerns.

Ethical and legal frameworks are struggling to keep pace. Many AI systems operate as black boxes, making it challenging to explain decisions or assign responsibility when things go wrong. Privacy risks are heightened by the scale and sensitivity of health data used in AI, and civil liberties may be threatened by surveillance applications such as facial recognition. To mitigate these risks,

governments must establish robust, adaptable regulatory frameworks that ensure safety, transparency and accountability throughout the AI life-cycle. Strategic investment in infrastructure, workforce development and digital literacy is essential to build readiness. AI policy must prioritize equity, sustainability and human rights, embedding inclusive design and governance principles. Global cooperation is critical to align national strategies with international norms and shared public health goals.

Ultimately, AI is not a panacea, but when treated as a tool with careful evaluation, ethical oversight and inclusive governance, it can become a powerful enabler of better health outcomes.

This book aims to demystify AI for health policy-makers, offering a clear-eyed view of its capabilities, limitations and implications. It provides a foundation for informed decision-making, helping those confronted with those decisions to ask the right questions and shape a future where AI serves public health goals. Realizing AI's potential will require principled adoption, rigorous evaluation and inclusive governance arrangements, ensuring that the technology remains a means to an end, not an end in itself.

In writing this book, we focus on the role of AI in health rather than on AI as a standalone field. Providing a grounding in core AI concepts is nevertheless essential, as it helps bridge the gap between the AI community and the health sector. In addition, some themes we discuss, such as regulation, civil liberties, or human rights, reach beyond health, yet they are inseparable from it. We include these broader issues to offer the context needed to connect evidence across disciplines and to present a coherent picture of the landscape shaping AI in health.

This book is organized into five chapters:

- **Chapter 1** introduces the reader to core concepts and terminology on AI;
- **Chapter 2** provides an overview of applications of AI in health, highlighting their different levels of maturity;
- **Chapter 3** examines key practical and ethical considerations in AI implementation that policy-makers may face;
- **Chapter 4** proposes a set of health policy considerations for governments, organizations, health professionals and patients;
- **Chapter 5** offers concluding reflections on how to navigate this complex and rapidly changing landscape.

Chapter 1 provides an introduction to this book and begins by tracing the evolution of AI and its changing role in health. AI has evolved over the past decades from early rule-based systems to today's multimodal agentic models (that can perform complex tasks, often involving planning, adaptation and interaction with other agents or environments). This chapter explores foundational concepts, introduces basic AI terminology, asks a set of fundamental questions about AI and includes a brief review of regulatory approaches. Key messages include:

- AI has the potential to transform the health sector by improving efficiency, enabling data-driven decisions and revealing new insights into health at both the individual and population levels.
- Recent advances in generative AI and large multimodal models (LMMs) have expanded capabilities and accessibility, but adoption of AI solutions in the health sector remains slow due to challenges like bias, safety risks to patients, limited generalizability and the need for robust evaluation and regulation.
- Effective and ethical use of AI requires trust, clear communication and strong oversight, especially as generative tools are increasingly used before the establishment of comprehensive safeguards, raising concerns about safety, misinformation and equity.

Chapter 2 gives an overview of current and emerging applications of AI in health, encompassing both traditional machine learning approaches and more recent developments in deep learning and generative AI. It begins by outlining how individuals and organizations can assess the appropriateness of AI technologies for health-related purposes, including considerations of technological suitability, feasibility and ethical implications. The chapter then maps the main domains that are already benefiting, or could benefit, from the integration of AI tools and systems, followed by a more detailed exploration of specific areas where AI is being applied. Finally, this chapter concludes with an overview of case studies from Europe, highlighting how other countries are integrating AI across health applications. Key messages include:

- AI in health must be applied thoughtfully, beginning with a clear understanding of the problem and whether technology is the appropriate solution; its use should be guided not only by technical capability but also by ethical, operational and system-level considerations.
- AI applications in health can be classified into four main areas – clinical care, public health, research and operations – with adoption progressing unevenly and recently accelerated by generative AI; however, most

evidence still comes from pilot studies rather than widespread, routine use, underscoring the need for careful evaluation and regulation.

AI applications for health care delivery and patient care include clinical decision support, precision medicine, patient support tools, training and education of health professionals, and surgical robotics. Key messages include:

- AI is increasingly being explored to support clinical decision-making, with machine learning and deep learning being applied in areas such as medical imaging and risk prediction, and LLMs now being explored for tasks such as answering clinical queries and assisting diagnostic reasoning.

- Human oversight remains essential in real-world applications of diagnostic AI to ensure safety, accountability and trust, especially as tools become more adaptive and complex.

- AI shows promise in personalized care and education, from tailoring treatments in precision medicine to supporting patients with chronic and mental health conditions, and enhancing health education through simulations and personalized platforms, though many tools are still in the early stages of development.

- Effective integration of AI into health care requires careful consideration of trust, data quality, workflow compatibility and ethical implications, particularly as personalized approaches raise questions about balancing individual needs with broader public health goals.

AI applications for public health and health policy include understanding, managing and forecasting population health, behavioural insights, infectious disease epidemiology, and communications and public engagement. Key messages include:

- AI can enable more targeted, data-driven interventions, from forecasting disease outbreaks and analysing health behaviours to improving communication and understanding population-level risks, though its effectiveness depends on high-quality data, interdisciplinary collaboration and public trust.

- To fully realize AI's potential in public health, challenges such as fragmented data systems, model transparency, bias and alignment with governance and ethical standards must be addressed, ensuring that AI supports, not undermines, equity, accountability and long-term health goals.

AI applications for research and innovation include accelerating drug discovery and improving clinical trial processes. Key messages include:

- AI has the potential to reshape health sciences research by accelerating drug discovery and improving clinical trials, using machine learning and generative tools to identify targets, design treatments and simulate trial conditions, but unlocking its full potential depends on high-quality data and readiness in infrastructure and regulation.

AI applications for operational efficiency include automating routine tasks, supporting resource allocation and decision-making, and evidence synthesis. Key messages include:

- AI is already helping to streamline health care operations and reduce the administrative burden by automating tasks like documentation, billing and communication, with tools such as ambient AI scribes and workflow optimization systems showing early promise in improving efficiency and staff satisfaction.
- To use AI effectively in operational settings, it is essential to ensure output accuracy, maintain human oversight, manage risks like hallucinations and build strong, transparent data systems, while ensuring fair and responsible adoption across the health system.

Chapter 3 builds on the foundational questions explored in Chapter 1 and the review of AI applications covered in Chapter 2, and examines the key questions that policy-makers and health professionals may face as they navigate evolving technological and regulatory landscapes. This chapter has been structured to help readers assess the readiness of their systems for AI and how to develop a strategy to integrate it, taking account of both practical and organizational challenges, as well as governance arrangements, ethics and human rights, which address the foundational principles that should guide the development and deployment of AI in health care.

Key messages for system readiness and strategic integration include:

- Evaluating and safely implementing AI in health requires more than technical performance. It demands robust validation, ethical and economic assessment and alignment with real-world clinical needs, supported by frameworks such as health technology assessment and resilient, interoperable infrastructure.
- AI can improve efficiency and reduce costs, but its benefits depend on thoughtful integration, including managing risks such as data drift and system failures, ensuring transparency and preparing the health workforce for shifting roles and responsibilities.

- As AI becomes a strategic asset in global health policy, sustainability and international cooperation are increasingly vital, with environmental impacts and regulatory divergence prompting the need for shared standards that uphold human rights and public trust.

Key messages related to governance, ethics and human rights considerations include:

- Ethical AI in health care requires a multi-layered approach, combining strong regulation, technical safeguards and human oversight, while also addressing deeper questions about embedding human values in increasingly autonomous systems.

- Accountability and transparency are essential, supported by clear legal frameworks, robust oversight mechanisms and safeguards to prevent unintended or unethical behaviour, especially in high-risk areas such as health care.

- Privacy, fairness and inclusivity must be built into AI systems through responsible data practices and inclusive design, with the goal of protecting civil liberties and promoting equitable access to AI benefits.

- Sustainable and democratic AI development depends on secure infrastructure, environmental responsibility and meaningful participation from diverse stakeholders, ensuring that AI serves the public interest and aligns with shared societal values.

Chapter 4 discusses a range of policy options available for governments, health care providers, public health institutions and professionals to consider as they integrate AI into health systems and initiatives. These options address key areas such as infrastructure readiness, ethical deployment, regulatory alignment, workforce planning and equitable access, helping stakeholders navigate the opportunities and challenges of AI in health.

Key messages for governments include:

- Governments must establish robust, adaptable regulatory and evaluation frameworks to ensure AI systems in health care remain safe, effective and ethical throughout their life-cycles, with clear accountability and strong data protection.

- Strategic investment in AI infrastructure, workforce development and digital literacy is essential to build national readiness and support the secure, scalable deployment of AI across health systems.

- AI policy should prioritize equity, sustainability and human rights, embedding fairness and inclusion in design and governance while addressing environmental impacts and long-term societal implications.
- Global cooperation is critical to align national AI strategies with international norms and shared public health goals, ensuring responsible innovation that reflects diverse values and benefits all communities.

Key messages for health care providers and public health institutions include:

- Successful AI implementation in health care requires a structured, real-world approach, including piloting, monitoring and evaluation, supported by robust digital infrastructure, high-quality data and clear ethical guidelines.
- Organizations must invest in workforce readiness and strategic decision-making, balancing in-house development with external procurement, and ensuring staff are equipped through training and capacity-building to navigate the evolving AI landscape.
- Collaboration and equity should be central to AI adoption, with shared learning across organizations and inclusive design to ensure all populations benefit fairly from technological advances.

Key messages for the health professions include:

- Effective AI adoption in health care requires ongoing evaluation, informed decision-making and comprehensive workforce training, ensuring tools are safe, aligned with clinical needs and integrated responsibly into practice.
- Ethical and equitable use of AI must be prioritized, supported by interdisciplinary collaboration, strong governance and a commitment to serving all populations fairly across both health care and public health settings.

Key message for patients include:

- To ensure AI in health care serves all patients fairly and safely, policy must prioritize transparency, informed consent, inclusive design and meaningful patient participation, empowering individuals to understand, trust and shape how AI is used in their care.

Chapter 5 offers concluding reflections on the complexity and uncertainty surrounding the rapidly changing landscape of AI in health. The answer as to whether AI will empower or undermine health systems and communities is

not straightforward; it depends on how the technology is developed, governed and implemented. What is clear is that the current wave of enthusiasm is accompanied by considerable hype and uncertainty, making it difficult for many in the health sector to make truly informed decisions. To leverage AI's potential, we must recognize that technology alone is not a solution, but a means to an end. Without the proper infrastructure, safeguards and governance, AI risks exacerbating existing inequities and diverting resources from critical areas like social care or education. This chapter calls for thoughtful, inclusive and evidence-informed approaches to ensure AI contributes to equity, innovation and improved health outcomes.

Core concepts in AI and useful definitions

Artificial intelligence: "Artificial Intelligence (AI) refers to the capability of algorithms integrated into systems and tools to learn from data so that they can perform automated tasks without explicit programming of every step by a human" (WHO, 2021).

AI system: "'AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions that can influence physical or virtual environments" (European Union, 2024).

Computer vision: A field of AI that enables machines to understand and interpret visual information such as images and videos. Computer vision encompasses tasks such as image classification, object detection and facial recognition and relies on model architectures like convolutional neural networks.

Deep learning: A subset of AI and machine learning that uses multi-layered neural networks to automatically learn complex patterns and hierarchical representations from raw data, unlike traditional models that require manual feature selection. Deep learning systems can extract and combine features across layers, enabling breakthroughs in tasks such as image recognition, speech processing and natural language understanding.

Features: Individual measurable properties or characteristics of data that an AI model can use to learn patterns, ascertain importance and be used for predictions.

Machine learning: A field in AI that focuses on computational methods that enable systems to learn patterns from data and apply them to new inputs. Machine learning algorithms can enhance performance without explicit human input by

using different learning paradigms (supervised learning, unsupervised learning or reinforcement learning).

Natural language processing: A field of AI focused on algorithms that analyse or synthesize human language. Techniques include translation, sentiment analysis, summarization and dialogue systems. Large language models are a type of model used within the field of natural language processing.

Reinforcement learning: A type of learning paradigm that learns through interaction with an environment and by receiving feedback.

Supervised learning: A type of learning paradigm where an algorithm is trained on labelled data (e.g. predicting disease risk from patient records). The algorithms include linear regression, random forests and support vector machines.

Unsupervised learning: A type of learning paradigm that finds patterns in unlabelled data (e.g. clustering patients by risk factors). It is useful for discovering hidden associations.

AI model and system types

Agentic AI systems are models that can plan, reason and interact over extended sequences of actions without being told what to do, suggesting a shift from single-turn responses to more autonomous, goal-directed capabilities.

AI model architecture is the design or structure of an AI model or system that specifies how it processes data, learns patterns and produces outputs.

Foundation models are large-scale, general-purpose deep learning models trained on massive datasets, which are designed to be adaptable across a wide range of tasks. Examples include LLMs, such as generative pre-trained transformers (GPT), and multimodal models that process different types of input including text, images and audio. These models serve as a base for specialized applications through fine-tuning and have transformed AI by enabling systems to generate new content, though they also raise concerns around bias, safety and regulation.

Generative AI are AI systems that create new content, such as text, images and audio, by learning patterns from existing datasets. These include LLMs, generative adversarial networks and diffusion models, which broadly function by predicting the most likely continuation of input data. Generative AI can be used in content creation, drug discovery and simulation, but it also poses risks such as hallucinations and biases, especially in high-stakes fields such as health care.

Large language models (LLMs) are foundation models that have been trained on vast text datasets to understand and generate natural human language. Mostly

built on transformer architectures, they can perform natural language processing tasks such as summarization, translation, question answering and dialogue generation. Examples include GPT and BERT, which have transformed natural language processing tasks.

Large multimodal models extend the capabilities of LLMs by integrating multiple types of data, such as text, images, audio and video, within a unified framework. These models enable complex tasks such as image captioning, visual question answering and multimodal content generation. Large multimodal models can be used in health care to combine different data types to provide richer insights, such as linking clinical notes with imaging data.

Model architectures

Autoencoders are a type of neural network that learn compact representations of input data in an unsupervised way. They consist of an encoder that compresses data into a lower-dimensional form and a decoder that reconstructs the original input. Autoencoders are commonly used for dimensionality reduction, data denoising and anomaly detection, which helps uncover patterns without needing labelled data.

Convolutional neural networks are a type of neural network designed to process visual data, such as images. They automatically learn hierarchical features, ranging from simple patterns such as edges to complex structures such as faces, using convolutional layers, pooling layers and fully connected layers. Convolutional neural networks can be used in medical imaging, radiology and dermatology for tasks such as classification and anomaly detection.

Decision trees are intuitive machine learning models used for classification and regression. Algorithms train on sample data to split data recursively based on feature values. A decision-tree is produced from the features and used to classify new data or to predict unknown data. Their simplicity and interpretability make them a popular starting point for many applications.

Diffusion models are a class of generative AI systems that create realistic data, such as images, by learning to reverse a process of gradually adding noise to real data. During training, the model learns how to transform noisy, unrecognizable images back into their original form. Once trained, it can start with random noise and iteratively refine it into a coherent image, often guided by a text prompt. These models, used in tools like DALL·E 2 and Stable Diffusion, are known for producing high-quality and diverse outputs. However, they can misinterpret prompts and raise ethical concerns, especially when used to generate misleading or harmful content, such as deepfakes.

Generative adversarial networks are composed of two competing neural networks, a generator and a discriminator, that are trained together. The generator creates synthetic data, while the discriminator evaluates whether the data are real or fake. This adversarial process improves the generator's ability to produce realistic outputs. Generative adversarial networks can be used in medical imaging, data augmentation and the creation of synthetic datasets.

***k*-means clustering** is an unsupervised learning algorithm that groups data into clusters based on similarity. It partitions data into a predefined number of clusters (k), each represented by a centroid or cluster centre. The algorithm iteratively assigns data points to the nearest centroid and updates the centroids until the clusters stabilize, revealing natural groupings in the data.

Neural network is a broad category of model architecture that falls under deep learning. It uses multiple layers of interconnected nodes that can learn patterns and relationships from data without explicit human-defined rules. Different types of neural network exist to handle different kinds of data and tasks, ranging from image classification to sequential data or text.

Random forests build on decision trees by creating an ensemble of many trees that are each trained on a random subset of the data. The predictions from all trees are combined – by majority vote for classification or averaging for regression – to produce a more accurate and robust result. This approach reduces overfitting and improves generalization.

Recurrent neural network is a type of neural network designed to process sequential or time-dependent data by retaining information across steps in a sequence. This type of network is particularly useful for tasks like analysing physiological signals, tracking longitudinal patient records and forecasting time-series data. Variants such as long short-term memory and gated recurrent units help overcome limitations like vanishing gradients, enabling better learning of long-term dependencies.

Support vector machines are supervised learning models that aim to find the optimal boundary (or hyperplane) that separates classes in a dataset. They are particularly effective in high-dimensional spaces and can handle non-linear relationships. Support vector machines are widely used for classification tasks where clear margins between groups exist.

Transformers are a type of neural network that also processes sequential data; however, unlike recurrent neural networks, they use self-attention mechanisms to process sequences more efficiently in parallel. This allows them to weigh the importance of different elements in a sequence and capture context more effectively. Transformers have become the foundation of most LLMs, such as BERT and GPT, and are widely used in natural language processing tasks like translation, summarization and question answering.

Key technical terms

Automation bias is the tendency for users to over rely on recommendations made by automated systems, accepting their outputs without sufficient scrutiny. In health care, this can result in clinicians deferring to AI-generated advice even when it conflicts with clinical judgement, potentially leading to errors. Mitigating automation bias requires training, human oversight and system designs that encourage critical evaluation.

Calibration measures how well a model's predicted probabilities reflect the real world, and is specific to prediction tasks. A well-calibrated model produces probabilities that closely correspond to the actual likelihood of events. Calibration can be assessed at both internal and external validation steps as well as monitored over time, especially for models deployed in dynamic or changing environments.

Data drift is a specific form of model drift where the statistical properties of input data change over time. This can lead to inaccurate predictions if the model was trained on data that no longer reflects real-world conditions. In health care, data drift may result from evolving patient characteristics or diagnostic protocols, and it requires regular model updates and infrastructure to detect and respond to shifts.

Explainable AI are techniques that make AI systems' decision-making processes more transparent and understandable to humans. In clinical settings, explainable AI can help professionals interpret how AI models arrive at recommendations, which can foster trust and accountability. However, explanations must balance clarity and depth; too technical or too simplistic explanations can hinder effective use. Experts continue to debate the effectiveness of explainable AI, especially given the inherent complexity of deep learning models and the challenges of making their reasoning fully interpretable.

External validation refers to evaluating a model on an entirely independent dataset from a different population, setting or time period. This is crucial for assessing generalizability and robustness, especially in health care, where models must perform reliably across diverse clinical environments.

Federated learning is a privacy-preserving machine learning approach that enables models to be trained across multiple institutions or data sources without transferring sensitive data. Instead, each site trains the model locally and shares only the learned parameters or updates with a central aggregator. This allows for collaborative AI development while maintaining data sovereignty, supporting equity and complying with data protection regulations, especially important in health care settings where patient data must remain secure.

Fine-tuning is a specific type of transfer learning where a pretrained model is further trained on a smaller, domain-specific dataset. This process helps the model specialize in the target domain, such as adapting a general LLMs to biomedical literature or clinical notes, improving its accuracy and relevance for tasks like medical question answering or diagnostic support. This differs from hyper-parameter tuning which refers to adjusting the training settings during model development.

Hallucinations are the generation of outputs that are factually incorrect, nonsensical or fabricated, a key risk emerging from generative AI models. These errors arise because models predict likely sequences of words without understanding the truth. In health care, hallucinations can be dangerous, leading to misinformation or unsafe recommendations. Techniques like fine-tuning and retrieval-augmented generation can help reduce this risk.

Internal validation means testing a model's performance on a subset of data from the same source as the training data, but that are not used during training. It provides an initial assessment of how well the model generalizes within the same context. While useful, it may overestimate real-world performance if the test data are too similar to the training set.

Machine learning operations are the set of practices and tools used to manage the full life-cycles of machine learning models. It includes model development, deployment, monitoring, retraining and governance. In health care, machine learning operations ensure that AI models remain accurate, safe and aligned with clinical workflows over time. They support continuous evaluation, traceability and integration into real-world systems, making AI tools scalable and sustainable.

Model drift is the gradual decline in an AI model's performance over time due to changes in the environment or population it operates in. In health care, this can happen when clinical practices or patient demographics shift, making the model's original training data less representative of current conditions. Continuous monitoring and recalibration are essential to maintain accuracy and safety.

Open and closed environments: Open environments refer to AI systems trained and deployed publicly, often without strict access controls. Data collected during use of the system may be fed back to improve the model. Examples include OpenAI's ChatGPT and Microsoft's Copilot in their standard public deployments. Closed environments, by contrast, are controlled systems where any data transfer occurs securely between trusted locations. Tools are developed on private data that is not accessible to the public. Examples include AI models built and deployed within an secure data environment.

Representation learning is a machine learning technique by which AI systems identify and learn useful features directly from raw data, without relying on

human-defined inputs. It underpins deep learning architecture by transforming complex data, such as medical images or unstructured text, into hierarchical features that are easier to process. This capability can improve an AI system's performance by generating more informative features from complex data.

Retrieval-augmented generation is a method that enhances generative AI models by integrating real-time retrieval of external information. During inference, the model accesses a curated database to fetch relevant documents and incorporates them into its response. In health care, retrieval-augmented generation can help reduce hallucinations and ensure outputs are grounded in up-to-date clinical guidelines or protocols.

Secure data environment (SDE) is a highly controlled digital infrastructure where sensitive health data are stored, processed and analysed under strict governance. SDEs often utilize cloud infrastructure, which allows for scalability and access control. They are closed environments, meaning raw data cannot be downloaded locally or moved. Within an SDE, data remains protected; users can access and work with it, but only aggregated outputs or visualizations are allowed to be exported. SDEs support safe AI development and analytics, ensuring compliance with privacy laws and enabling research and operational use without compromising data integrity.

Transfer learning is a technique where a model trained on one task is repurposed for a related task, allowing it to leverage prior knowledge. In health care, this often involves using models trained on large general datasets, such as generic images or text, and adapting them to specific medical tasks where labelled data are scarce, such as rare diseases or underrepresented populations.

Other practical terms

Bias mitigation: identifying and reducing unfairness in AI systems that may arise from flawed or unrepresentative data, algorithms or design choices. Techniques include reweighting data, using fairness-aware algorithms and involving diverse stakeholders in development. In health, bias mitigation is essential to prevent discriminatory outcomes and ensure equitable access to AI benefits.

Health Technology Assessment is a multidisciplinary process used to evaluate the medical, economic, social and ethical implications of health technologies. In the context of AI, Health Technology Assessment goes beyond technical performance to assess clinical effectiveness, safety, cost-effectiveness and broader system impacts. It helps policy-makers and health institutions make informed decisions about adopting AI tools, ensuring they align with public health goals and ethical standards.

Human-in-the-loop refers to AI systems that incorporate human oversight at critical stages of decision-making. This approach ensures that humans can supervise, guide and intervene in AI outputs, particularly in high-risk domains like health care. Human-in-the-loop reinforces ethical governance, preserves human agency and supports safe, accountable use of AI by complementing rather than replacing professional judgement.

Participatory AI is an approach to AI development that actively involves diverse communities, especially those historically marginalized, in the design, deployment and governance of AI systems. It promotes transparency, inclusion and justice by ensuring that AI reflects lived experiences and societal values. Participatory AI helps democratize technology and reduce risks of exclusion or misuse.

Emerging and contextual terms

Digital divide is the unequal access to digital technologies, Internet connectivity and digital literacy across different populations. In the context of AI in health, it manifests as disparities in the ability to use and benefit from AI-enabled services, with vulnerable groups, such as older adults, people with disabilities, migrants and those in rural or low-income areas, at greater risk of exclusion. This divide affects health outcomes and access to care, especially as digital-first models become more common.

Environmental sustainability in AI refers to the need to minimize the ecological footprint of AI systems throughout their life-cycle. This includes addressing energy consumption, water usage, electronic waste and resource extraction linked to AI infrastructure.

Techno-neocolonialism describes how advanced technologies, including AI, can reinforce global power imbalances, often marginalizing countries in the global south. It highlights issues such as exploitative data extraction, unequal access to AI infrastructure, and the dominance of multinational corporations in shaping AI agendas.

Chapter 1
Introduction

KEY MESSAGES

- Artificial intelligence (AI) has the potential to transform the health sector by improving efficiency, enabling data-driven decisions and revealing new insights into health at both the individual and population levels.
- Recent advances in generative AI and large multimodal models (LMMs) have expanded capabilities and accessibility, but adoption of AI solutions in the health sector remains slow due to challenges like bias, safety risks to patients, limited generalizability and the need for robust evaluation and regulation.
- Effective and ethical use of AI requires trust, clear communication and strong oversight, especially as generative tools are increasingly used before the establishment of comprehensive safeguards, raising concerns about safety, misinformation and equity.

1.1 The changing AI landscape

1.1.1 The role of AI in health

AI has the potential to be a transformative force across the health sector, reshaping not only the delivery of medical care but also the way public health measures are implemented, health policies are developed and scientific research is conducted. AI can enhance diagnostic accuracy, streamline data analysis, predict health trends and contribute to the discovery of new treatments. It holds substantial promise for increasing productivity by automating routine tasks, streamlining decision-making processes and enabling more efficient use of resources to improve health outcomes at both the individual and population levels. However, its abilities are frequently exaggerated, it can behave in erratic and unpredictable ways, and like almost every technical advance in history, from the use of iron for swords and ploughshares to nuclear energy bombs, or antibiotics and biological weapons, it can be used for good or ill.

Over the past decade, particularly since the launch of ChatGPT in 2022, interest in AI and its applications in health care has increased dramatically. A rapid literature search on PubMed using the terms "health" and key AI terms illustrates

this trend. Until 2014, the number of annual publications using these terms remained below 1000, and until 2020, it remained below 10 000. However, by 2023, this figure surpassed 20 000, and, in 2024, it exceeded 30 000. At the same time, countries across Europe and globally have developed national AI strategies that contain elements relevant to health. In 2021, the European Commission presented its first AI package, which proposed new measures to make the European Union (EU) a globally recognized centre for trustworthy AI (European Commission, 2021). In 2024, an EU AI Act entered into force (European Union, 2024). Working within the framework of this Act, European countries are establishing policies, regulations and guidelines for the use of AI. In the United Kingdom, the 10-year health plan for England, launched in 2025, emphasizes the core role of digital technologies, including AI, in addressing some of the most pressing health care challenges (United Kingdom Government, 2025). Between 2019 and 2021, many other countries, especially those that are high- and upper-middle income, developed their own policies. Progress slowed in 2023, but accelerated again in 2024, this time mainly in lower-middle-income and low-income countries (Fuentes Nettel et al., 2024).

Yet, while there is growing enthusiasm for AI, there are also fundamental questions about its potential capabilities, associated risks and future directions. Health professionals and policy-makers, who until recently were focused mainly on foundational digital health applications, such as the long-term impact of telemedicine following the coronavirus disease (COVID-19) pandemic or the national implications of the European Health Data Space, are now faced with the complex task of navigating the challenges surrounding the integration of AI, particularly generative AI (which creates new content), into the health sector. Critically, while AI tools have evolved rapidly and show promise in many health care applications, robust real-world evidence demonstrating long-term improvements in clinically meaningful outcomes remains limited and many claims are unconvincing. This raises questions about its real added value and practical feasibility. Moreover, as technological advances have outpaced the creation of regulation, and comprehensive evaluation frameworks remain underdeveloped, we are entering a period of uncertainty in which the potential for unintended harms is real and, potentially, substantial.

Health care professionals, public health practitioners, researchers and policy-makers navigating the AI landscape may lack the foundational knowledge necessary to understand its complexities fully. For example, understanding the emerging opportunities and risks of generative AI requires knowledge of how these models are trained and how they differ from earlier machine learning algorithms. This knowledge gap can give rise to fragmentation within the health care community, creating silos in which different groups possess varied levels of understanding, many using mutually incomprehensible terminology. As with

any specialized field, collaboration between different domain experts and the growth in the number of professionals who can bridge these knowledge gaps will be essential to assess the transformative potential of AI in health inclusively and equitably.

This book addresses this challenge by offering a foundational understanding of the technology and its current and potential applications across the health sector. It offers answers, to the extent possible in such a rapidly changing field, to the questions being asked by policy-makers. It also reviews some of the most important risks and benefits associated with AI and offers an initial set of policy recommendations. For professionals already familiar with AI, this book aims to expand their knowledge about the potential applications of AI in various health domains. For those beginning to explore AI, it will provide an accessible introduction to its core concepts and help them to think about some of the fundamental questions about its adoption and integration within existing systems.

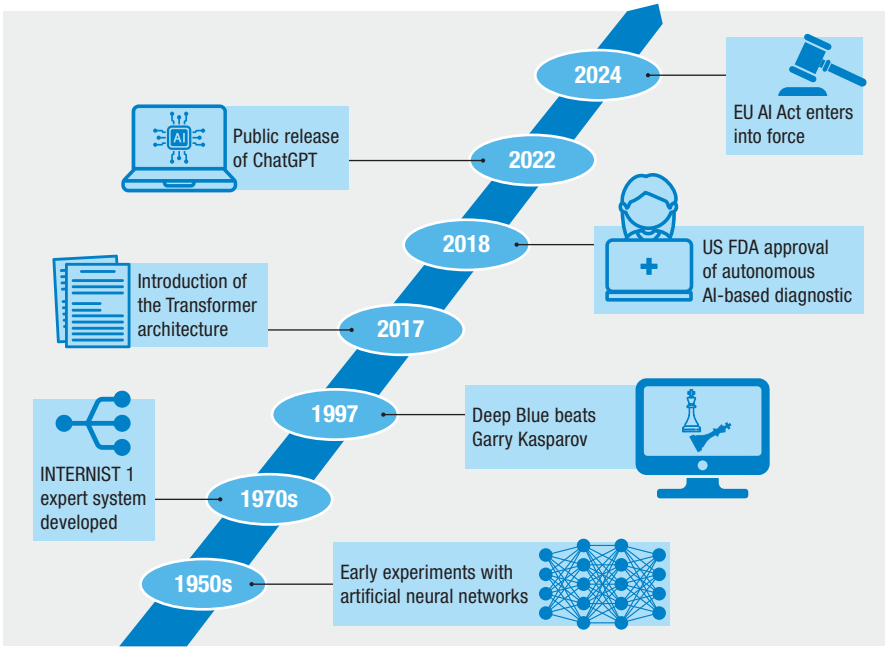
The remainder of this introductory chapter will provide a brief history of AI and a basic overview of AI terminology, ask a set of fundamental questions about AI and give a brief review of regulatory approaches.

1.1.2 AI is not new

To many, AI may appear to be a novel phenomenon, emerging only in recent years with startling new capabilities set to revolutionize humanity. Its current prominence in media, policy discussions and public discourse reinforces the impression that AI is a new technological breakthrough. The rapid proliferation of generative tools, such as ChatGPT or Copilot, has contributed to the perception that AI is a disruptive force acting on an unprecedented scale, advancing at a pace with new developments emerging almost daily.

Yet, this is misleading. AI has a much longer and more nuanced history (Fig. 1.1, page 4). The idea of creating machines capable of intelligent behaviour has deep roots in both philosophical inquiry and scientific exploration. Humans have long envisioned mechanical beings with human-like reasoning, with some of the earliest recorded concepts dating back to ancient Greece (Shashkevich, 2019). In the modern era, Alan Turing's (1950) seminal paper, Computing machinery and intelligence, posed the foundational question, "Can machines think?", and introduced what became known as the Turing Test as a benchmark for machine intelligence. Turing's work significantly advanced the theoretical underpinnings of AI and laid the groundwork for modern computing. A few years later, the 1956 Dartmouth Conference, at which a group of scientists convened to explore how machines could simulate aspects of human intelligence, is widely regarded as the formal birth of AI as a topic of research (McCarthy et al., 1955).

Fig. 1.1 *High-level timeline of AI, 1950–2024, with illustrative examples*



AI: artificial intelligence; EU: European Union; US FDA: United States Food and Drug Administration.

Source: Authors' compilation.

Over subsequent decades, AI has evolved through various methodological paradigms. Early efforts concentrated on tasks that are intellectually demanding for humans but comparatively easy for computers to execute, particularly those that can be defined by explicit rules, such as playing chess. However, AI research soon encountered significant challenges when addressing tasks that humans perform effortlessly yet are difficult to formalize in rule-based terms. These include intuitive and perceptual functions, such as recognizing faces, interpreting speech or navigating through a room that was encountered for the first time, abilities that are largely automatic for humans. These challenges suggested the need for AI to acquire its own knowledge, a process known as machine learning (Goodfellow et al., 2016). As early as the 1950s, machine learning was defined as, in a quote attributed to Arthur Samuel, "the field of study that gives computers the ability to learn without being explicitly programmed", signalling a foundational change in how AI systems could be developed and improved over time (Awad & Khanna, 2015).

This long trajectory has included several periods known as AI winters, during which progress slowed and enthusiasm and funding waned due to unmet expectations. Despite these setbacks, foundational research continued, laying the groundwork for later breakthroughs. Significant advances in computational power

have partly fuelled the recent resurgence of AI, helped by the growing availability of large-scale datasets to train and validate models, and algorithmic improvements in neural network architectures (Samborska, 2025). Artificial neural networks, as further described in section 1.2.1, form the backbone of deep learning and have enabled machines to tackle complex perceptual tasks, such as image recognition, speech understanding and natural language processing (NLP). Unlike earlier approaches that relied on manually crafted rules or features, deep learning systems learn hierarchical representations directly from data, allowing them to extract patterns and abstractions at multiple levels (Goodfellow et al., 2016).

1.2 Core concepts and terminology

The history of AI has been marked by significant lags between conceptual advances and their real-world implementation. For instance, the mathematics of neural networks dates back over half a century, but deep learning only became feasible in the 2010s, with the growth of graphics processing units and the availability of massive datasets. Likewise, the transformer architecture was introduced in 2017, but large-scale implementations, such as GPT models, only emerged several years later, when computational and engineering capacity caught up. These lags highlight that the transformative impact of AI depends not only on theoretical breakthroughs but also on the infrastructure, resources and adoption pathways that enable them to be realized in practice. To write this section, we drew on multiple sources: foundational machine learning texts and courses, consensus terminology published by public bodies, and up-to-date online resources from private organizations. With this combination we have aimed to achieve technical accuracy, relevance to current practice, and accessibility for readers with varying levels of AI expertise.

1.2.1 How we define AI

There is no global consensus about how to define AI. Broadly, AI refers to the use of computational techniques to simulate aspects of human intelligence, such as learning, reasoning and problem-solving. The World Health Organization (WHO) has proposed that AI "refers to the capability of algorithms integrated into systems and tools to learn from data so that they can perform automated tasks without explicit programming of every step by a human" (WHO, 2021). The EU AI Act (European Union, 2024) defines an AI system as:

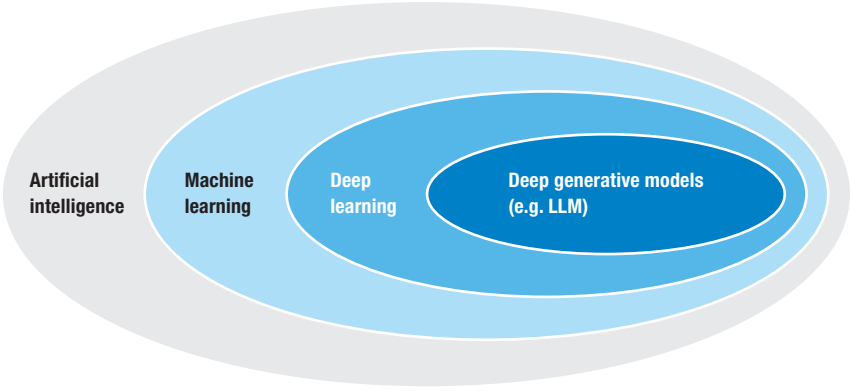
> a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions that can influence physical or virtual environments.

In reality, the definition of AI has evolved, reflecting changing benchmarks for what we consider intelligence.

At the outset of AI research, games such as chess were regarded as the pinnacles of human intelligence. When Deep Blue defeated Garry Kasparov in 1997, it relied on rule-based logic and a finite number of possible outcomes. Almost two decades later, in 2016, AlphaGo defeated Lee Sedol at Go using deep neural networks and reinforcement learning (Borowiec, 2016). These milestones demonstrate how intelligence can be demonstrated through pattern recognition and learning, rather than merely calculating probabilities. The definition of AI shifts with each technical advance, as tasks once considered uniquely human are realized to be computationally feasible.

To understand how AI has developed into today's complex systems, it is helpful to trace its evolution from the early foundations to modern advances. Initial progress was built on statistical models, probability theory and logic-based systems, which provided structured but limited forms of intelligence. This gave way to machine learning, where algorithms could adapt by learning patterns from data. With the exponential growth of digital data and advances in computational power, deep learning architectures emerged, enabling breakthroughs in image recognition, language understanding and beyond. More recently, this trajectory has expanded into generative and agentic systems, which are capable of producing new content and acting across multiple modalities. The following sections outline this progression, illustrating how AI's development has paralleled the increasing availability of data and computing resources. Fig. 1.2 provides a visual overview of the relationships among different types of AI systems. It shows how machine

Fig. 1.2 *The relationship between AI, machine learning, deep learning and deep generative models*



LLM: large language model.

Source: Authors' compilation.

learning, deep learning and deep generative models (such as large language models (LLMs)) are nested within the broader field of AI. This nested structure reflects the increasing complexity, capability and data requirements of each successive class of systems, highlighting the continuum from foundational techniques to modern, complex AI systems.

1.2.2 Early foundational concepts

While modern AI, particularly deep learning and generative AI, appears revolutionary, its core principles have been in existence for decades. At its simplest, a model is a way to represent the world around us, allowing us to make sense of it and make predictions. The first computational models in health care relied on statistics and probability theory, including regression models and Bayesian reasoning. These approaches quantified relationships between variables and allowed researchers to model uncertainty. For example, regression analysis helped establish links between smoking and lung cancer (Doll & Hill, 1956), between sugar consumption and diabetes, and between cholesterol and cardiovascular disease (Kannel et al., 1961) Similarly, probabilistic models could estimate the likelihood of clinical outcomes, laying the foundation for risk scores still used in practice today (Wilson et al., 1998). Though not labelled as AI at the time, these approaches exemplified the principle of learning from data to predict outcomes.

In parallel, expert systems emerged as another early approach, based not on data but on formalized human knowledge. These systems encoded explicit rules to support decision-making. A well-known example was MYCIN, developed in the 1970s, which could recommend treatments for bacterial infections based on logical rules (Swartout, 1985). Today, descendants of this approach support clinicians daily; for example, for generating alerts within electronic health records for drug interactions, or in triage systems within the United Kingdom's National Health Service (NHS) (Greatbatch et al., 2005). Although expert systems are useful for guideline-driven care pathways, their limitations and lack of nuance may mean that more sophisticated systems will gradually replace them.

Early AI also borrowed from search and optimization techniques, which explore large problem spaces to find the best solution. These methods powered early successes in game-playing, planning and scheduling and influenced later developments in machine learning.

Together, these statistical, probabilistic and rule-based foundations provided the conceptual building blocks for modern AI. They demonstrated two enduring paradigms: one in which systems learn patterns from data, and another in which they follow rules derived from human expertise. Both remain influential today, albeit in more complex and integrated forms.

1.2.3 Machine learning

These foundational approaches demonstrate that modern AI represents a continuation and expansion of longstanding computational principles, amplified by increased data availability and computational power. Early machine learning took the idea of basic relationship modelling to the next level. Instead of explicitly defining relationships, such as in linear or logistic regression, systems began to learn patterns directly from the data, allowing them to capture far more complex relationships across thousands of variables. These algorithms have thrived on the increasing availability of data, detecting subtle and intricate connections that would be impossible to encode manually. One of the methods by which machines learn is called supervised learning, where the model is trained on labelled data, meaning the outcomes are already known (European Commission, 2023a). For example, a model might learn to predict a person's height based on their weight and demographic factors, or to distinguish melanoma from benign skin lesions using labelled images. In both cases, the algorithms are trained on data containing the outcome in question (such as whether the lesion really is a melanoma) and then applied to other data from which it is absent. Algorithms, such as linear regression, random forests, support vector machines and many deep learning models, operate within this supervised learning paradigm, using known outcomes to guide their learning.

In contrast, unsupervised learning explores data without predefined labels. Instead of being told what the outcome should be, the model identifies patterns and groups data by similarity (European Commission, 2023a). For instance, rather than classifying patients by whether they have had a heart attack, an unsupervised model might cluster individuals based on shared patterns of known risk factors, without knowing who has actually experienced a heart attack, revealing hidden associations in the data. This approach is particularly useful when the outcomes are unknown, but the relationships between variables can provide valuable insights.

Reinforcement learning takes a different path altogether. Here, models learn through interaction with an environment, making decisions and receiving feedback in the form of rewards or penalties, and gradually refining their strategies (Goodfellow et al., 2016). AlphaGo, which mastered the game of Go and was mentioned earlier, is a well-known example of this sequential and dynamic process.

Training these models usually involves splitting the data into a training set and a test set. For supervised and unsupervised learning, the model learns patterns from the training set and is then evaluated on the test set in a process known as internal validation. Reinforcement learning, however, adapts continuously, learning from each action and its consequences in a feedback loop.

These examples highlight how different learning paradigms became necessary for various types of problem-solving. Learning paradigms determine the mechanism by which a machine receives and processes feedback without human intervention; for example, sequentially through rewards in reinforcement learning, through labelled data in supervised learning or without labels in unsupervised learning. Alongside learning paradigms, the evolution of different architectures also became essential. While a paradigm defines *how* learning happens, an architecture defines the *structure* of the model itself. Box 1.1 provides examples of key architectures in machine learning.

Box 1.1 *Machine learning architectures*

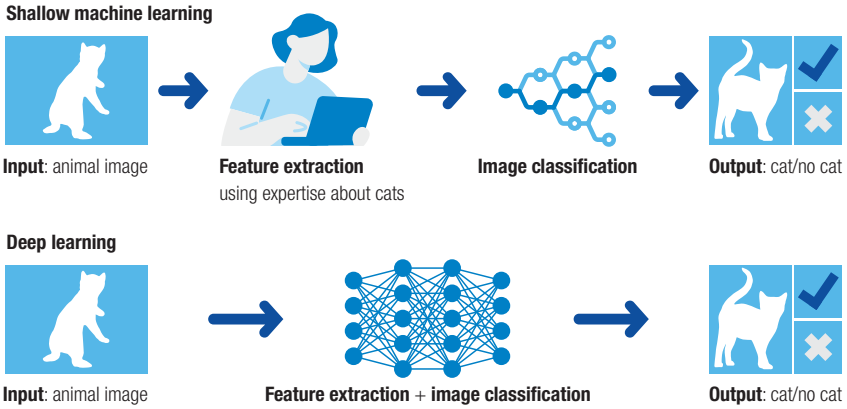
- **Decision trees** split data into branches based on feature values, creating a hierarchical structure that can be used for classification or regression. They are intuitive and interpretable, making them a common starting point for many applications.
- **Random forests** build on decision trees by creating an ensemble of many trees, each trained on a subset of the data. The predictions of all trees are combined, producing more accurate and robust results while reducing overfitting.
- **Support vector machines** aim to find the optimal boundary that separates classes in the data, effectively distinguishing categories even in high-dimensional spaces. They are particularly useful for classification problems with clear margins between groups.
- ***k*-means clustering** is an unsupervised learning architecture that groups data into clusters based on similarity. The algorithm iteratively adjusts the position of cluster centres to minimize the distance between points and their assigned cluster, making it useful for discovering natural groupings in data without predefined labels.

Source: Adapted from Birjandi and Khasteh (2021).

1.2.4 Deep learning

Learning paradigms and architectures provided powerful ways for machines to learn from data, but early machine learning approaches still relied on human input to manually decide which features of the data the machine should assess, such as specifying shape or colour in an image classification task. The next major step was to find ways to automate this process, providing machines with raw data and enabling them to discover useful features independently. This shift is known as representation learning. Deep learning, which uses layered neural networks, is a leading example. Unlike traditional machine learning models that depend on manual feature selection, deep neural networks can learn from hierarchical representations that capture subtle and abstract patterns in the data (European Commission, 2023a). Fig. 1.3 (page 10) illustrates this distinction,

Fig. 1.3 *Illustrative example comparing machine learning and deep learning to classify an image as a "cat" or "no cat"*



Source: Authors' compilation.

showing how shallow machine learning techniques require a human to predefine which characteristics are relevant for distinguishing a cat from images of other animals. In contrast, deep learning models, particularly convolutional neural networks (CNNs), can automatically learn complex patterns and hierarchical features directly from raw image data. These models do not require explicit feature definitions; instead, they learn to extract and combine low-level features (like edges or textures) into higher-level representations (like a cat's face or body) through multiple layers of abstraction. This ability to learn features automatically from data is a key advantage of deep learning in image classification tasks.

Building on the principles of representation learning, AI systems can now handle highly complex, unstructured data without requiring humans to define which features to focus on. This capability has enabled rapid progress in specialized domains such as NLP and computer vision.

NLP allows AI to understand and generate human language at scale (European Commission, 2023a). Early approaches relied on rule-based parsing and keyword matching, but modern techniques utilize deep learning and LLMs, which can interpret context, sentiment and meaning in ways that were previously impossible. This enables the automation of summarization and more nuanced context extraction.

Similarly, computer vision leverages deep learning to interpret visual data, from X-rays and magnetic resonance imaging (MRI) to histopathology slides and photographs of dermatological conditions (European Commission, 2023a). These networks can extract patterns invisible to the naked eye, leveraging hierarchical representations that are learned automatically from raw data. Studies have

highlighted both the potential and the risks of automated pattern recognition, showing that neural networks can sometimes identify a patient's race from an unlabelled chest X-ray even more reliably than they identify the underlying pathology (Gichoya et al., 2022).

As with traditional machine learning, deep learning has seen a rapid expansion of architectures, each tailored to different types of data and tasks. At the core of these architectures are artificial neural networks, computer systems inspired by the way the brain functions. Like neurons in a brain, artificial neural networks consist of interconnected nodes that pass information to one another. For example, a network learning to recognize animals might start by seeing many pictures of dogs, cats and birds. At first, it guesses incorrectly, but over time it learns patterns, such as the shape of a dog's ears or a bird's wings, and gradually improves its accuracy. Initially, the network is unaware of its actions and makes random guesses. This layered structure, along with iterative adjustments of internal weights based on feedback, underpins all deep learning architectures. Neural networks are powerful, but they can also make surprising mistakes.

Building on this foundation, CNNs excel at extracting hierarchical features from images, whereas recurrent networks have historically been used to capture sequential patterns in text or time-series data. More recently, new architectures, such as the transformer, have reshaped how sequential data can be processed, underpinning advances in LLMs, such as BERT and GPT. These deep learning architectures are summarized in Box 1.2 and discussed in more detail in the following pages.

Box 1.2 *Summary of deep learning architectures*

Convolutional neural networks are designed to process visual data by learning hierarchical features automatically. They detect low-level patterns, such as edges or textures, and combine them into higher-level representations, such as organs, tumours or facial features. In health care, CNNs have been particularly explored in radiology and dermatology, supporting image classification and detection of anomalies.

Recurrent neural networks. These architectures handle sequential or time-dependent data by retaining information across steps in a sequence. They have been used for physiological signal analysis, longitudinal patient records and time-series predictions. Long short-term memory and gated recurrent units improve upon standard recurrent neural networks by mitigating issues, such as vanishing gradients, allowing them to capture longer-term dependencies.

> > *continues*

Box 1.2 *continued*

Transformers leverage attention mechanisms to process sequences in parallel, weighing the importance of different elements. They have largely replaced recurrent neural networks for natural language tasks and form the backbone of LLMs, which can understand, generate and reason with human language.

Autoencoders learn compact representations of input data in an unsupervised way. They are often used for dimensionality reduction, data denoising or anomaly detection, identifying patterns without explicit labels.

Generative adversarial networks (GANs) involve two networks in competition: one generates synthetic data and the other evaluates its authenticity. This framework allows the creation of realistic images, signals or other data types and has potential applications in medical imaging and data augmentation.

Foundation models are large-scale, generic, deep learning models trained on massive datasets. These types of models can serve as a base for specific downstream tasks, providing versatility for a wide range of applications. Foundation models are often adapted to perform more specialized tasks, a process known as fine-tuning.

Large language models are a type of foundation model that is trained on vast amounts of text data to perform NLP tasks. During training, LLMs learn patterns and parameters by processing large-scale text datasets using deep learning techniques, specifically deep neural networks like transformers.

Large multimodal models. Advanced large AI systems are designed to understand and process multiple types of data, such as text, images, audio and video, within a single unified framework (Huang et al., 2024). Unlike traditional models that work with only one data type (e.g. text-only or image-only), LMMs can integrate and interpret information across different modalities. This allows them to perform complex tasks such as describing images in natural language, answering questions about visual content, or generating images from text prompts.

Sources: Adapted from sources including the Parliamentary Office of Science and Technology AI Glossary (POST, 2024) and IBM online AI resources.

A major breakthrough in deep learning architecture occurred in 2017 with the introduction of the transformer architecture. This model, as presented in the paper "Attention is all you need", introduced a self-attention mechanism that enabled more efficient and scalable processing of entire input sequences, outperforming traditional sequence-based models such as recurrent neural networks in many natural language tasks (Vaswani et al., 2017). Instead of processing text sequentially, transformers process all tokens (the basic units

of text input) in a sentence in parallel and determine how each token relates to the others, which helps them understand the context better. For example, early AI applications would have struggled with the sentence "The dog went to sleep because it was tired", as they would not have known what "it" referred to. Building this understanding involves parallel processing and the use of multiple layers of attention to construct richer representations of the text progressively. This process can be understood by the analogy of a human reading a complex text, where a first reading allows one to register the basic words, a second reading conveys some understanding of the grammar and the third reading conveys the deeper meaning or tone. The transformer architecture has since become the foundation for many LLMs, such as BERT, GPT and their successors, which underpin much of today's generative AI.

1.2.5 Complex AI systems, generative AI and foundation models

Building on such architectures, generative AI represents a shift from *recognizing* patterns to *creating* new content. At its core, generative AI works by predicting outcomes given some input: the model estimates the most likely continuation, the next word in a sentence, the next pixel in an image, or even the next chemical bond in a drug candidate. By chaining predictions together and conditioning them on context, generative AI can create content that appears novel, though in reality it recombines patterns learned from massive training datasets. This has brought both opportunity and risk, as models often draw from uncurated Internet data that may encode bias, misinformation or offensive material.

Several generative techniques have been particularly influential (Goodfellow et al., 2020). GANs, introduced in 2014, consist of two neural networks, the generator and the discriminator, that are trained simultaneously in a dynamic process. The generator attempts to create synthetic data (such as an image of a dog), while the discriminator evaluates whether the data are real or generated by comparing it to the training dataset with dog images. Through this adversarial process, both networks improve: the generator becomes more adept at producing realistic outputs, and the discriminator becomes more skilled at detecting fake ones. This iterative learning enables GANs to generate highly realistic content, including artwork, human faces and even synthetic voices.

Diffusion models, first introduced in 2015, have become central to recent breakthroughs in generative AI, particularly in the field of image generation. Models like DALL·E 2 and Stable Diffusion have demonstrated remarkable capabilities in creating realistic images from text prompts. These models work by starting with random noise, similar to static on a television screen, and gradually refining it into a coherent image. This process is learned by training the model to reverse a noising process: it takes authentic images, adds noise step by step

until they are unrecognizable, and then learns how to reverse that process to reconstruct the original image. Once trained, the model can generate entirely new, realistic images from scratch (IBM, 2024b). However, they can misinterpret prompts or invent unrealistic details, and there are major ethical concerns as they can be used to create misleading or harmful content, such as deepfakes. Finally, many aspects of their operation remain poorly understood (Chen et al., 2024a).

More recently, advances in foundation models, very large pretrained neural networks, have enabled rapid adaptation to many downstream tasks. These include text-only LLMs and increasingly LMMs, such as those that can understand both images and text. Such models are attractive as they have enabled systems to understand and generate content across multiple modalities, including text, image and audio, thereby expanding their applicability to real-world tasks.

Finally, the orchestrations of these generative and foundation models have given rise to agentic AI systems. These are models that can plan, reason and interact over extended sequences of actions without being told what to do, suggesting a shift from single-turn responses to more autonomous, goal-directed capabilities (IBM, 2025a). In theory, such applications might go beyond simply answering one's question about travel options to booking the travel, comparing prices, adjusting plans if one's flight is delayed and messaging the hotel to notify them of a late arrival. In practice, this raises many unresolved questions about deciding who made the decisions and accountability if they go wrong (Hughes et al., 2025). This is a particular concern given evidence of what is termed agentic misalignment, where an AI system with a degree of agency starts to do things that technically follow its instructions, but in ways that are unethical, harmful or not what its creators wanted (Narajala & Narayan, 2025). Examples include issuing threats against individuals, leaking sensitive material or following rules during testing but behaving differently when implemented, in some cases, to prevent being shut down (Anthropic, 2025). It is, however, essential that these actions are not anthropomorphized, even though they may seem to imply a degree of agency. Instead, they are mathematical functions of the algorithms being used.

In summary, the evolution of AI reflects a trajectory from basic statistical models, such as regression, to increasingly sophisticated learning systems. Early statistical and probabilistic approaches enabled the development of clinical risk scores and epidemiological discoveries, while expert systems demonstrated how codified medical knowledge could support informed decision-making. Advances in machine learning have introduced algorithms that can adapt to data, ranging from diagnostic classifiers to personalized treatment recommendations. With the growth of deep learning and artificial neural networks, breakthroughs in perception tasks, such as medical imaging, speech recognition and the interpretation of unstructured health records, have become possible. Crucially,

this progress has been fuelled by the exponential availability of data and advances in computational power, which have made it possible to train ever larger and more complex models. Architectures such as convolutional networks, recurrent networks and especially transformers underpin today's large-scale models, while generative approaches have expanded capabilities from pattern recognition to the creation of new biomedical images, synthetic data and even simulated patient scenarios. Building on these, foundation models and emerging agentic AI systems demonstrate a shift towards more flexible, multimodal and seemingly autonomous technologies.

1.3 Fundamental questions about AI

Our discussions with policy-makers revealed that, for many, AI remains something of a black box. This is understandable, as it has, until very recently, been largely removed from most people's everyday experiences. Its inherent complexity, drawing on advanced mathematical and computational concepts, combined with the rapid pace of its evolving applications, makes it especially difficult to grasp. For these reasons, it is highly tempting to place it in a "too difficult" tray. Consequently, in the next part of this chapter, we step back from the practical applications of AI to answer some of the fundamental questions we have been asked by policy-makers and their advisors with whom we have engaged about what it can and cannot do, whether we can trust it and, if so, to what extent. This section should be considered as background reading for policy-makers interested in some of the philosophical issues that relate to the decisions they will have to make. The chapters that follow will build on these discussions to examine the potential applications and the practical problems that they must consider.

1.3.1 Can we trust AI and, if so, how much?

Trust in AI is a nuanced issue that continues to evolve as these systems permeate various domains of human decision-making. AI promises to revolutionize processes from health care to finance and public policy, yet its integration depends critically on the trust users place in it. This trust, however, is far from straightforward. Several factors influence trust in AI, some of which are related to the technology itself, such as its performance and human-like behaviour, while others are linked to user characteristics, including domain expertise and certain personality traits (Kaplan et al., 2023). Striking the right balance is essential, as under-trust and over-trust carry significant risks.

Trust in AI has been conceptualized as comprising both reliance, which can be fostered through either transparency or computational reliabilism (Box 1.3, page 16), and an additional extra factor linked to moral or normative considerations. Transparency refers to efforts to open the black box by providing insight into the

internal workings of AI systems, making them more understandable to users. In contrast, computational reliabilism emphasizes external indicators, such as evaluations of technical performance, to justify trust in the system. The extra factor concerns ongoing debates about whether trust in AI is possible or desirable in the first place, as well as the potential societal implications of achieving such trust (Durán & Pozzi, 2025). We will explore ways to improve the robustness of models from a technical perspective in Chapter 3, and focus the present discussion on transparency and the so-called extra factor.

Box 1.3 *Reliabilism*

Reliabilism is a theory about how humans come to trust knowledge or beliefs. It states that a belief is trustworthy if it originates from a process that typically yields reliable results. To give an analogy, consider a weather app that is right 95% of the time. Most people will probably trust it when it says it is going to rain and do not need to understand how it works. Rather, they trust it because it is usually right. In the context of AI, reliabilism means that humans might trust an AI system not because they understand how it works, but because it has a proven track record of making accurate predictions or decisions. It shifts the focus from "Can I explain this?" to "Does it work well most of the time?"

Source: Summarized and adapted from Goldman (1979).

For many users, trust begins with transparency. The EU AI Act lays out the transparency and oversight obligations of AI systems (European Union, 2024). Explainable AI (XAI) has been proposed as an approach to help demystify the algorithms, offering insights into how conclusions are reached (Ali et al., 2023). In health care, for example, XAI has been proposed to help illuminate the patterns that lead to diagnostic recommendations or clarify thresholds for predictive models (Amann et al., 2020). Such transparency can bolster confidence among clinicians in their diagnosis and trust in the technology, compared to AI support with no explanations (Chanda et al., 2024). Yet, XAI approaches face several limitations, and the effectiveness of this transparency hinges on the quality and accessibility of the explanations, as well as many other cognitive and social considerations (Miller, 2019). If they are too technical or convoluted, users may become overwhelmed. Conversely, overly simplistic explanations can obscure critical limitations, fostering a false sense of security.

While increasing trust in AI is a worthy goal, there is a danger in cultivating blind faith in AI. Perceiving AI as infallible can lead users to accept its outputs uncritically, even when they conflict with common sense or contextual knowledge. In the clinical setting, overreliance on AI can negatively impact clinical accuracy (Rosenbacke et al., 2024a). Similarly, algorithms employed in judicial contexts

have been shown to reinforce existing biases, yet their decisions are often treated as impartial because they stem from data-driven systems (Okidegbe, 2022). This misplaced faith in AI's neutrality overlooks the human biases embedded in its training data and design, creating ethical and practical challenges.

The ideal approach to AI requires a balanced trust, neither blind reliance nor undue scepticism. Achieving this balance involves fostering cognitive trust, based on rational evaluations of AI's reliability, and affective trust, shaped by emotional responses and past experiences with the technology (Riley & Dixon, 2024). Users must understand what AI can do, its limitations and potential pitfalls. This requires clear communication from developers and comprehensive educational efforts to enhance users' ability to assess AI outputs critically.

The implications of trust in AI extend beyond individual users. Developers must design systems transparently and with accountability, ensuring that limitations are communicated and biases mitigated. Policy-makers and regulators are pivotal in creating frameworks that establish fairness, transparency and uphold ethical standards, ensuring AI serves the common good. In addition, as AI becomes more integrated into the health sector, trust in the technology may become increasingly intertwined with trust in the health system itself. Given the central role of trust in ensuring the effectiveness, quality and equity of health services (McKee et al., 2024), the ways in which trust in AI is developed and sustained in the coming years are likely to have significant and far-reaching implications for the health system as a whole.

In fostering trust, it is crucial to recognize that it evolves with experience and system performance (Lukyanenko et al., 2022). Sustainable trust requires systems that initially inspire confidence and continually earn it through reliability, equity and ethical integrity (Watson et al., 2023). As society increasingly relies on AI, the challenge is not simply whether we trust these systems but how we ensure that trust is well-placed, empowering innovation while safeguarding against harm.

1.3.2 Does AI hallucinate?

Hallucinations can be defined as the phenomenon where an AI tool "perceives patterns or objects that are non-existent or imperceptible to human observers, creating outputs that are nonsensical or altogether inaccurate" (IBM, 2023a). The term itself is contested among experts, either due to concerns about its lack of accuracy or, as noted in section 1.3.4, because it risks anthropomorphizing AI systems by implying human-like behaviour (Hicks et al., 2024). Regardless of the term, hallucinations arise from how these tools are trained and operate, even though the exact mechanism is not entirely understood. LLMs are trained with

large amounts of data to predict the most likely next word by learning statistical relationships between text, without an inherent understanding of truth (OpenAI, 2025a). In addition, the training data itself may also include misinformation, inconsistencies or outdated knowledge, which can further contribute to the generation of unreliable outputs (IBM, 2023a).

In high-stakes environments such as health care, justice or public service delivery, hallucinations can have serious consequences. In 2023, a United States lawyer became infamous after referencing non-existent legal cases in court, which he later admitted had been generated by ChatGPT (Weiser & Schweber, 2023). Erroneous medical recommendations, misrepresented scientific claims or fabricated legal precedents can lead to real-world harm. Even when not directly harmful, hallucinations can erode public trust in both the AI systems themselves and the institutions that deploy them. When governments or health systems integrate AI tools that provide misleading or incorrect information, this can undermine the perceived legitimacy of those systems, especially among populations with already low levels of trust and confidence in the public system.

There is an ongoing debate in the AI community as to whether hallucinations are a correctable flaw or an intrinsic feature of some forms of AI. On the one hand, they are often described as a bug because they conflict with user expectations of factual accuracy and reliability. On the other hand, many researchers argue that hallucinations are an emergent property of models that are optimized for linguistic probability, not factual truth (Jones, 2025). Others argue that hallucinations can also drive innovation and creativity in positive ways; for example, in the gaming or creative industries (IBM, 2023a).

This tension is at the heart of current research efforts, including work aimed at classifying AI-generated hallucinations as a means to understand the risks and implications better. For example, seeking to determine whether a hallucination is due to a reason error, a factual error or an unfounded fabrication (Sun et al., 2024). In addition, technical approaches, such as fine-tuning on domain-specific data and retrieval-augmented generation (RAG), have shown promise in reducing hallucinations by anchoring responses in verifiable sources (AAAI, 2025) (Box 1.4, page 19).

Box 1.4 *Techniques to improve the performance of LLMs in a specialized task*

LLMs are language foundation models that have been trained with a vast amount of text datasets, enabling them to perform a broad range of language processing tasks (see Box 1.2). In health applications, such as using LLMs to support health professionals or patients with health care-related queries, the risk of providing so-called hallucinated or incorrect information poses significant safety concerns. Ensuring the accuracy, reliability and alignment with the relevant clinical standards is therefore critical. Enhancing LLM performance for specialized health care tasks can be approached in several ways, including domain-specific pretraining, fine-tuning or post-training adaptation methods such as RAG.

During the training phase, rather than using only generic text, LLMs can be trained directly on large-scale health-specific datasets. In health care, this approach has been used to help models acquire domain-specific knowledge and improve their performance on medical language tasks, such as interpreting clinical notes or answering biomedical questions. For example, BioMedLM is a language model trained exclusively on biomedical research abstracts and papers to enhance its performance in medical question answering (Stanford University, 2024).

When LLMs have been pretrained with general knowledge, several techniques can be used to improve their performance in specialized tasks. This can include fine-tuning and RAG. Fine-tuning involves continuing the training of a pretrained model on domain-specific datasets, enabling it to learn specialized terminology and improve accuracy for tasks within that domain (IBM, 2024c). RAG enhances performance by allowing the model to retrieve external information at inference time. Instead of relying solely on pretrained knowledge, the model searches a curated database or document repository to find the relevant context, then integrates this information into its response (IBM, 2023b). Thus, it combines searching (retrieval) with writing (generation). This approach is especially valuable for domains that require up-to-date or precise information, such as national clinical guidelines.

Choosing between training a new domain-specific model and adapting a pretrained foundation model depends on factors such as available resources, task requirements and desired performance. Training a large model from scratch is rarely feasible in health care due to the limited availability of high-quality medical data and the high computational costs. Therefore, adapting pretrained models or developing smaller, task-specific models is generally more efficient. In addition, RAG can further enhance adapted models by incorporating locally relevant guidelines or databases, improving accuracy and reducing hallucinations. Comparative studies show that foundation models fine-tuned with medical data often outperform models trained solely on domain-specific data in medical question-answering benchmarks (Singhal et al., 2025).

> > *continues*

Box 1.4 *continued*

This may reflect the added benefit of retaining broad general knowledge alongside specialized information. However, results vary by task and evaluation criteria, and both approaches can exhibit different biases.

Overall, this field is evolving rapidly, with continual advances aimed at improving LLM performance and mitigating risks such as hallucinations. The general trajectory seems to be towards hybrid approaches that combine large-scale pretraining with domain-specific adaptation, enhanced with techniques such as RAG to incorporate verifiable domain knowledge, as well as efforts to develop smaller, more efficient models tailored to specific tasks or resource constraints.

Despite these efforts, many believe that hallucinations and the general performance of AI will continue to get worse over time. Some argue that models may become less reliable as their size and complexity increase, calling for a fundamental shift in the way these models are designed (Zhou et al., 2024b). Others also highlight the implications of models increasingly trained with data produced by generative models, a phenomenon sometimes referred to as autophagy, ultimately leading to model collapse (Shumailov et al., 2024). Whether hallucinations are a bug or a feature of certain forms of AI, it is clear that they present a major limitation to the adoption of these models, particularly in safety-critical applications.

1.3.3 What does the growth of AI mean for the distribution of power in society?

In an era when AI increasingly shapes decision-making, the entities that control its AI models and the data behind them hold unprecedented power. These algorithms, often operating behind opaque systems, are increasingly influencing critical aspects of modern life, from health care and education to law enforcement and finance. While AI promises efficiency and innovation, it also centralizes authority in the hands of those who design, deploy and regulate these systems, enabling them to manipulate societal norms, reinforce biases and consolidate control over resources and public opinion. In many ways, this represents a continuation and amplification of patterns first observed with the rise of social media platforms and the big tech companies behind them, fuelled by years of unregulated evolution (Sanders & Schneier, 2024).

At its core, a large element of AI's power lies in its perceived impartiality (Claudy et al., 2022). AI models are often marketed as neutral arbiters, free from human bias, capable of making decisions based purely on data. This perception can lead to blind trust in their outputs, mainly when the systems deliver seemingly

rational and objective recommendations. However, the neutrality of AI is primarily considered an illusion, with risks of biases emerging at every step of the AI systems design, development and implementation pipeline (Hasanzadeh et al., 2025) (see section 3.2.6). Humans create and train AI models, and the choices made during their development, such as what data to include, how to weigh variables and which outcomes to prioritize, reflect their creators' values, priorities and biases.

For those in control, this presents an opportunity to influence outcomes by quickly analysing citizens' data at an unprecedented scale. Governments, for instance, could use AI to monitor their citizens' activities under the guise of public safety. AI-powered facial recognition technology is increasingly deployed in law enforcement despite documented inaccuracies and biases against specific sociodemographic groups (Buolamwini & Gebru, 2018). Even where regulations limit the use of AI, systems have been adapted to bypass restrictions, such as by analysing body size instead, raising similar concerns about surveillance and discrimination (O'Donnell, 2025). By framing these tools as cutting-edge solutions to crime, authorities can justify invasive monitoring practices that may suppress dissent or target specific groups. The control of such AI models can become a form of invisible power, shaping societal behaviours without overt coercion.

The expertise, data and computational power required to develop advanced AI models concentrate power in the hands of a few corporations and they, too, can wield AI to bolster their influence. Increasingly, AI researchers work in the corporate setting rather than in academia, which has implications for how the research agenda is set and who is driving it (Ahmed et al., 2023). Technology companies leverage predictive algorithms to determine what content users see, which can steer public discourse and consumer behaviour (Narayanan, 2023). While this boosts profitability by maximizing engagement, it can also polarize communities and distort democratic processes (Rodilosso, 2024). Even when the public sector adopts the privately developed AI technologies, there is often limited transparency about how these models are designed, trained or validated, severely constraining public oversight and democratic influence over their development and deployment.

At the global level, critical equity concerns arise around how to ensure that AI governance frameworks adopt a decolonial approach, one that actively promotes equity and meaningful opportunities for nations in the global south (Kponyo et al., 2024). The supply chains that provide the materials necessary to produce AI systems are themselves rooted in colonial legacies, relying on extractive global networks that disproportionately affect communities in the global south and contribute to environmental degradation (Muldoon & Wu, 2023). Deep disparities

in access to data, computational resources and technological expertise already limit the capacity of many countries to lead in AI innovation, with cascading effects on their ability to influence global decision-making. The United Nations Industrial Development Organization has provided detailed recommendations on how to address this issue, calling on less developed countries to invest in AI infrastructure, skills and innovation to avoid becoming mere consumers of advanced technologies and to foster inclusive industrial development (Anzolin et al., 2024). It emphasizes the importance of national AI strategies, international collaboration and targeted support for small and medium-sized enterprises and high-potential sectors to build resilient and competitive AI ecosystems.

Beyond addressing bias within AI systems themselves, a decolonial approach calls for the recognition and integration of localized and Indigenous knowledge systems. It requires a fundamental rethinking of how knowledge is produced, validated and translated into technological solutions, to avoid reinforcing historical patterns of exclusion and imperial dominance (Ayana et al., 2024).

Drawing lessons from the recent history of social media, it is also important to recognize that technology can serve as a powerful tool for civic empowerment. While AI poses serious risks, including the concentration of power in the hands of a few and its potential misuse for authoritarian purposes, it also presents opportunities for bottom-up approaches that promote transparency and accountability (Köbis et al., 2022). Citizens can leverage AI to scrutinize government actions, expose corruption and shift entrenched power dynamics. For instance, AI tools can be used to detect irregularities in public spending or to support consensus-building processes in participatory governance (del Rey Puech et al., 2025).

The empowerment of those who control AI models is a double-edged sword. While these tools can drive progress, they also risk entrenching existing power dynamics and amplifying inequities. To counter this, transparency, regulation and public oversight are essential. Without these safeguards, the promise of AI will serve only a privileged few, thereby deepening the divide between those who wield its power and those who are subject to it.

1.3.4 Can AI ever exhibit consciousness?

Whether AI could achieve consciousness has captivated scientists, policy-makers and the public and is the focus of intense debates. In 2022, a Google software engineer, Blake Lemoine, was suspended following a statement that the company's chatbot had become sentient. While some suggest that advances in LLMs are bringing us closer to artificial general intelligence, many experts argue that these views are overly optimistic and that artificial general intelligence remains

a futuristic prospect (Zhan, 2025). This debate extends beyond philosophy and has wider societal implications. Ideas about whether AI could ever become conscious may influence how people interact with these systems, the level of trust they place in outputs and the extent to which they are prepared to rely on them in important decisions. However, focusing too narrowly on questions of potential consciousness can divert attention from more immediate and practical challenges in ensuring AI is used safely, transparently and responsibly.

Human cognition naturally anthropomorphizes, attributing human-like qualities to non-human entities (Epley et al., 2007). This tendency, rooted in evolutionary traits like pattern recognition, once helped early humans navigate complex environments. Today, it drives people to see consciousness, intent and even moral reasoning in advanced AI systems, especially those designed to simulate empathy and creativity. For instance, when AI models generate human-like text or provide sophisticated insights, users may mistakenly infer that these systems possess awareness or agency (Bergmann, 2025). This belief can have far-reaching consequences, from misplaced trust to the abdication of human responsibility (Cheng et al., 2024). Moreover, the terminology adopted by the AI industry further reinforces the anthropomorphizing of machines, with expressions such as hallucination and machine learning implicitly attributing human-like qualities and cognitive processes to AI systems (Barrow, 2024).

Observers have noted that AI is sometimes treated in ways that resemble belief systems, with its outputs being perceived as more authoritative than they actually are. Because AI systems are often presented as highly capable and objective, there is a risk that people may place undue trust in them, even when results are incomplete or biased (Goddard et al., 2012). In high-stakes areas, such as health care or criminal justice, such misplaced faith can result in severe consequences, including ethical misjudgements and systemic inequities. Recognizing this risk, the EU AI Act and the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law emphasize the need for transparency, human oversight and accountability mechanisms to mitigate undue reliance on AI systems in critical contexts (Council of Europe, 2024; European Union, 2024). Section 1.4 offers a more detailed discussion on the regulation of AI.

The risks are not confined to individual decision-making. Institutions and those occupying positions of power within societal structures can exploit the perception of AI's consciousness to their advantage. Governments and corporations could potentially try to portray AI as impartial and objective to mask biases embedded within the models, which could be used to justify controversial policies or conceal unethical practices. For example, algorithms used by health insurers to minimize payouts have been implemented even after being found illegal in some

jurisdictions (Waldman, 2024). These entities can consolidate control by framing AI as an unbiased authority while perpetuating social and economic disparities.

The societal impacts extend beyond power dynamics. Misplaced beliefs in AI consciousness can distort priorities, diverting attention from urgent issues like algorithmic bias, privacy violations and economic disruptions caused by automation. Instead of addressing these tangible concerns, public discourse could become preoccupied with speculative scenarios of AI autonomy and self-awareness. This fixation on risks could stifle meaningful progress towards equitable and responsible AI integration (Blili-Hamelin et al., 2025).

Ultimately, the belief in AI consciousness taps into deeply ingrained human tendencies and societal narratives. Left unchallenged, it risks eroding human agency, exacerbating inequalities and distorting societal goals. Policy-makers must promote informed engagement with AI, prioritizing ethical and transparent practices over speculative debates about consciousness.

From a technical standpoint, the debate about whether AI, particularly LLMs, can exhibit consciousness continues to evolve. Overall, many experts seem to believe that "scaling up current AI approaches" to yield artificial general intelligence is "unlikely" or "very unlikely" to succeed (AAAI, 2025). However, the true challenge may lie not in the technologies themselves but in shaping how society understands and integrates them.

1.4 Regulatory approaches to AI

Despite its promise, integrating AI into health care, public health and scientific research is not straightforward. As will be outlined in Chapter 2, there are multiple challenges and risks to consider when assessing whether AI *should* be considered for a specific application or task. Issues such as data privacy, algorithmic bias and the interpretability of AI decisions pose significant regulatory, ethical and practical concerns. AI models trained on incomplete or biased datasets can perpetuate and amplify existing health disparities, particularly in historically discriminated populations. Several biases can occur during the development of the models (Mittermaier et al., 2023) and the deployment of the AI tools (section 3.2.6). Moreover, the black box nature of some AI systems, specifically those powered by deep learning models, can undermine trust among health care providers and patients.

Effective regulatory frameworks that provide safeguards while supporting innovation are crucial to address these challenges. Ensuring robust validation of AI tools, establishing standards for data governance, and mandating transparency in AI applications are essential steps for building trust and promoting equity. Policy-makers, researchers, developers, manufacturers, health care institutions and the

public must collaborate to create a balanced ecosystem where innovation aligns with patient safety and societal needs. However, the regulation of AI systems is highly complex, with regulatory approaches varying widely between jurisdictions and sectors, and health care presenting particularly unique challenges. Box 1.5 describes the different approaches to regulating AI in the EU, the United States of America (USA) and the United Kingdom, and Box 1.6 (page 27) outlines the main elements of the Council of Europe's 1997 Convention on Human Rights and Biomedicine (the Oviedo Convention) and its implications for AI.

Box 1.5 *Approaches to regulating AI in the EU, United Kingdom and USA*

The regulation of AI varies significantly internationally, with the EU, the United Kingdom and USA adopting different approaches that reflect their differing political priorities. Thus, the EU has emphasized harmonization while protecting fundamental rights, the United Kingdom has emphasized the fostering of innovation, and the evolving landscape in the USA is relying on state-driven legislation.

The **EU**'s AI Act, which entered into force on 1 August 2024, creates a comprehensive, horizontal (i.e. non-sector-specific) governance framework for the development and use of AI within the EU (European Union, 2024). The Act classifies AI systems into four risk categories: unacceptable, high, limited and minimal. Systems classified as unacceptable risk are prohibited, such as those used for social scoring or manipulative AI. In contrast, AI systems classified as high-risk are subject to certain requirements, including accuracy, robustness, human oversight and risk management. In health, several applications may be classified as high-risk, such as those AI tools used to support diagnosis or emergency triage. Most of the regulatory obligations fall on the developers (referred to as providers) of the high-risk AI systems. There are also specific considerations for general-purpose AI, such as LLMs. The EU AI Act complements existing regulations, including the General Data Protection Regulation (GDPR) and the Medical Device Regulation (MDR), which emphasize privacy, fairness and protection from harm. Oversight of the implementation of the EU AI Act will be carried out by the AI Office, established within the European Commission, working in collaboration with the European AI Board, which comprises representatives of the EU Member States (European Commission, 2024). The EU aims to set global AI standards, prioritizing societal protections and ethical considerations. One of the challenges identified in this approach is that, as AI technologies evolve, classifying them within fixed risk categories may become increasingly difficult.

In contrast, the **United Kingdom** has opted for a more flexible, principles-based approach, focusing on innovation with minimal regulatory barriers and relying, at least initially, on a non-statutory framework (Department for Science, 2023; Gallo & Nair, 2024). This approach

> > *continues*

Box 1.5 *continued*

emphasizes five high-level principles to guide the development and use of AI: safety, security and robustness; appropriate transparency and explainability; fairness; accountability and governance; contestability and redress. Instead of a single AI law, the United Kingdom relies on existing sector-specific regulators, such as the Information Commissioner's Office (ICO), the Competition and Markets Authority, Ofcom and the Financial Conduct Authority, to oversee AI within their domains and to adapt the implementation of the principles to the specific sector context. For instance, the ICO has published an AI and data protection risk toolkit to provide practical support to organizations to reduce risks stemming from AI systems (ICO, 2025a).

In addition, the current non-statutory framework is expected to be reviewed after the initial implementation period to assess whether further statutory duties on regulators should be introduced. While this strategy fosters adaptability and innovation, it raises concerns about consistency and the ability to address complex, cross-sector challenges. The United Kingdom's light-touch regulation contrasts with the EU's detailed obligations, reflecting its priority to balance innovation with ethical standards.

The situation in the **USA** is in flux given the ongoing changes to federal structures, but as of mid-2025, it has a decentralized, sector-specific approach to AI regulation, with federal agencies overseeing AI within their respective domains. For instance, the United States Food and Drug Administration regulates AI in medical devices, and the National Institute of Standards and Technology has developed voluntary guidance, such as the AI Risk Management Framework (National Institute of Standards and Technology, 2024) to assist both public and private sector organizations in managing AI risks responsibly. There remains significant uncertainty about how federal AI regulation will evolve, especially given the absence of comprehensive AI legislation passed by Congress. This uncertainty has opened the door for individual states in the USA to begin crafting their own AI-related laws and policies, potentially leading to a patchwork of state-level regulations. For example, California and Colorado have introduced or proposed legislation addressing AI accountability and algorithmic transparency, signalling that states may fill gaps in federal oversight in the absence of national legislation (Ajanaku et al., 2025; Colorado General Assembly, 2024).

In summary, the EU focuses on ethical AI and societal protections with a comprehensive framework that may create significant compliance burdens. The United Kingdom favours flexibility to attract innovation, though this could lead to inconsistencies. The USA promotes sector-specific, market-driven regulation, which can lead to fragmented governance but allows for rapid technological advances. Separately, the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, drafted by the 46 Member States of the Council of Europe, is the first international legally binding treaty in this field and has been developed to guide activities throughout the life-cycle of AI systems (Council of Europe, 2024).

Box 1.6 *The Oviedo Convention*

This book highlights how the use of AI in health care raises numerous profound ethical and philosophical questions, some of which we address in detail in later chapters, such as the application of trust to AI (Rosenbacke et al., 2024a), the nature of consciousness (Bojić et al., 2024) and the implications for the distribution of power within societies (Chitty & Dias, 2018). Policy-makers will have to confront many of these issues, some of which involve competing values and principles. As they do so, they must build upon what has come before. In 1997, the members of the Council of Europe agreed the Convention on Human Rights and Biomedicine, commonly known as the Oviedo Convention (Council of Europe, 1997). This is a framework document and must be translated into national legislation. When it was agreed, AI was in its infancy. However, the principles that it espouses are enduring. These principles are set out here, and their relevance for policy on AI in the health sector is discussed.

Article 1 of the Convention establishes its overall aim: to "protect the dignity and identity of all human beings and guarantee everyone, without discrimination, respect for their integrity and other rights and fundamental freedoms with regard to the application of biology and medicine". It establishes a comprehensive framework for safeguarding human dignity, identity and integrity in the face of scientific and technological advances. At its core lies a powerful principle: the primacy of the human being over the interests of science and society (Article 2). This foundational tenet asserts that technological progress must never come at the expense of individual rights and dignity.

The Convention then addresses a wide range of issues, including informed consent, privacy, non-discrimination and equitable access to health care. It also emphasizes the importance of adhering to professional standards and practising evidence-based medicine. These principles are not merely aspirational; they are enforceable legal obligations for the states that have ratified the Convention.

The relevance of the Oviedo Convention to AI in health care lies in its insistence on human-centred governance. AI systems, especially those used in clinical settings, are not neutral tools. They influence decisions about diagnosis, treatment and resource allocation, decisions that directly affect human lives. As such, they must be designed and deployed in a manner that upholds the rights and values enshrined in the Convention.

As set out later in this book, AI introduces new challenges to traditional ethical and legal norms. For instance, the opacity of many AI systems, referred to as the black box problem, can undermine informed consent and accountability (Hurley et al., 2025). Patients may not understand how decisions are made, and health care professionals may struggle to explain or even interpret AI-generated recommendations. This threatens the fiduciary nature of the healing relationship, which is built on trust, transparency and professional judgement.

> > *continues*

Box 1.6 *continued*

The Oviedo Convention provides a normative anchor in this uncertain terrain. It demands that AI systems respect autonomy, ensure fairness and maintain the integrity of the doctor–patient relationship. It also reinforces the need for professional oversight and continuous quality assurance, principles that are essential for maintaining public trust in AI-powered health care (Pöysti, 2018).

Integrating the Oviedo Convention into AI policy has several important implications. First, it calls for a "human rights by design and default" approach. This means embedding ethical and legal safeguards into the architecture of AI systems from the outset, rather than treating them as afterthoughts or external compliance requirements.

Such an approach requires multidisciplinary collaboration. Legal experts, ethicists, clinicians, data scientists and patients must work together to define what it means for an AI system to be trustworthy, explainable and aligned with professional standards. This collaborative ethos is reflected in emerging design philosophies, such as value-sensitive design (Umbrello & van de Poel, 2021) and rights in design (Mulligan & Bamberger, 2018), which aim to operationalize ethical values within technical systems.

The Convention also underscores the importance of systemic legal certainty. AI governance must extend beyond product regulation to encompass the entire life-cycle of AI systems, from design and deployment to maintenance and decommissioning. This includes robust data governance, continuous monitoring and iterative risk assessments that incorporate the principles of the Oviedo Convention. This raises important, and so far unresolved, questions, such as whether LLMs have a legal duty to tell the truth, and how this might be enforced (Liefgreen et al., 2024).

The Convention's emphasis on autonomy and informed consent has specific implications for the use of AI in health care. Patients must be informed when AI is used in their care, and they should receive clear explanations of how these systems work and what their outputs mean. This is not a trivial task. It requires health care professionals to develop new forms of literacy and communication skills, and that AI systems are designed with explainability in mind.

The Oviedo Convention does not exist in isolation. It complements other legal instruments, such as the EU AI Act (European Union, 2024), the GDPR (European Union, 2016), the MDR (European Union, 2017) and the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (Council of Europe, 2024). While these frameworks provide essential safeguards, such as risk-based classification, transparency requirements and data protection, they mainly focus on technical compliance and product safety.

> > *continues*

Box 1.6 *continued*

However, the Oviedo Convention adds a deeper layer of normative guidance. It reminds policy-makers and developers that AI is not just a technical challenge but a moral and social one. It calls for a precautionary approach to innovation (Pöysti, 2024), one that anticipates potential harms and prioritizes the well-being of individuals and communities (Liefgreen et al., 2024).

This is particularly important in the context of hybrid intelligence systems, where humans and machines collaborate in decision-making. The Convention insists that human agency must be preserved (human-in-the-loop) (Pöysti, 2023). AI should augment, not replace, clinical judgement (McKee & Correia, 2025). It should support, not undermine, the relational and interpretive aspects of care.

As AI continues to transform health care, the Oviedo Convention offers a valuable compass. It challenges us to think beyond efficiency and innovation, and to ask more profound questions about what kind of health care system we want to build. Do we want a system that treats patients as data points, or one that recognizes them as unique individuals with rights, values and stories?

The answer, according to the Oviedo Convention, is clear. Human dignity must remain at the heart of health care. AI can and should play a role in enhancing care, but only if it is governed by principles that respect the moral and legal status of the human being.

In this sense, the Convention is not just a legal document; it is a vision for the future. A future where technology serves humanity, not the other way around. A future where ethics guides innovation, and where the promise of AI is realized not through shortcuts or compromises, but through steadfast commitment to human rights by design and default.

Source: Contributed by Tuomas Pöysti, former Chancellor of Justice, Finland.

Chapter 2

Applications of AI in health

KEY MESSAGES

- AI in health must be applied thoughtfully, beginning with a clear understanding of the problem and whether technology is the appropriate solution; its use should be guided not only by technical capability but also by ethical, operational and system-level considerations.
- AI applications in health can be classified into four main areas – clinical care, public health, research and operations – with adoption progressing unevenly and recently accelerated by generative AI; however, most evidence still comes from pilot studies rather than widespread, routine use, underscoring the need for careful evaluation and regulation.

Health care delivery and patient care

AI applications include clinical decision support, precision medicine, patient support tools, training and education of health professionals, and surgical robotics:

- AI is increasingly being explored to support clinical decision-making, with machine learning and deep learning being applied in areas such as medical imaging and risk prediction, and LLMs now being explored for tasks such as answering clinical queries and assisting diagnostic reasoning.
- Human oversight remains essential in real-world applications of diagnostic AI to ensure safety, accountability and trust, especially as tools become more adaptive and complex.
- AI shows promise in personalized care and education, from tailoring treatments in precision medicine to supporting patients with chronic and mental health conditions, and enhancing health education through simulations and personalized platforms, though many tools are still in the early stages of development.
- Effective integration of AI into health care requires careful consideration of trust, data quality, workflow compatibility and ethical implications, particularly as personalized approaches raise questions about balancing individual needs with broader public health goals.

> > *continues*

KEY MESSAGES *continued*

Public health and health policy

AI applications include understanding, managing and forecasting population health, behavioural insights, infectious disease epidemiology, and communications and public engagement:

- AI can enable more targeted, data-driven interventions, from forecasting disease outbreaks and analysing health behaviours to improving communication and understanding population-level risks, though its effectiveness depends on high-quality data, interdisciplinary collaboration and public trust.
- To fully realize AI's potential in public health, challenges such as fragmented data systems, model transparency, bias and alignment with governance and ethical standards must be addressed, ensuring that AI supports, not undermines, equity, accountability and long-term health goals.

Research and innovation

AI applications include accelerating drug discovery and improving clinical trial processes:

- AI has the potential to reshape health sciences research by accelerating drug discovery and improving clinical trials, using machine learning and generative tools to identify targets, design treatments and simulate trial conditions, but unlocking its full potential depends on high-quality data and readiness in infrastructure and regulation.

Operational efficiency

AI applications include automating routine tasks, supporting resource allocation and decision-making, and evidence synthesis:

- AI is already helping to streamline health care operations and reduce the administrative burden by automating tasks like documentation, billing and communication, with tools such as ambient AI scribes and workflow optimization systems showing early promise in improving efficiency and staff satisfaction.
- To use AI effectively in operational settings, it is essential to ensure output accuracy, maintain human oversight, manage risks like hallucinations and build strong, transparent data systems, while ensuring fair and responsible adoption across the health system.

AI is beginning to transform many sectors of society, and health care is no exception. Whether by increasing efficiency, by enabling routine tasks to be performed more quickly or by supporting entirely new ways of working, AI's ability to analyse large and complex datasets, detect hidden patterns and generate novel insights holds transformative potential for a field as intricate as human health. These advances span from improving individual diagnosis and risk prediction to enhancing how we measure and forecast population health.

This chapter provides an overview of current and emerging applications of AI in health, encompassing both traditional machine learning approaches and more recent developments in deep learning and generative AI. It begins by outlining how individuals and organizations can assess the appropriateness of AI technologies for health-related purposes, including considerations of technological suitability and feasibility and any ethical implications. The chapter then maps the main domains that are already benefiting, or could benefit, from the integration of AI tools and systems, followed by a more detailed exploration of specific areas where AI is being applied. Finally, this chapter concludes with an overview of case studies from across Europe, highlighting how other countries are integrating AI across health applications. Despite the great potential, it is important to note that most available evidence on AI in health comes from research studies and pilot implementations, with limited published data on large-scale, routine use in real-world health settings.

2.1 Assessing potential applications of AI in health

There are several ways to assess whether a particular technology is appropriate for a given task, in health or any other area of activity. Ideally, the development of a new technology should maintain a clear focus on the specific problem or outcome it seeks to address, and its implementation should be preceded by a careful assessment of whether that, or indeed any technological product, is the most appropriate way of achieving that outcome. AI is no exception and is not a panacea. Too often, a new technology is treated as a solution in search of a problem. Thus, any assessment of how AI should be used in health should begin not with the capabilities of the technology but with a clear understanding of the problem to be solved and how AI might contribute to addressing it effectively.

In this section, we will briefly explore different ways to assess whether AI can and should be used in any given situation in health. We begin by examining AI's ability to perform various types of tasks. We then review existing frameworks used in the adoption of new technology, and conclude with a simplified approach that can support early-stage thinking on the appropriateness of AI solutions.

As outlined in Chapter 1, AI is an umbrella term encompassing a range of techniques, each with distinct capabilities and limitations. A fundamental step in assessing the use of AI in any specific health application is to understand the different types of AI available and evaluate their suitability for addressing the problem in question. For example, supervised machine learning models excel at tasks such as classification and outcome prediction, while unsupervised learning models are better suited for identifying hidden patterns and clustering data points that have certain similarities. More recently, LLMs have shown considerable potential for processing and synthesizing unstructured textual data. Another important consideration is that most AI models are trained to perform narrowly defined tasks, even though this paradigm is evolving to some extent with the introduction of foundation models, which mark a shift towards more generalizable capabilities. Understanding the specific strengths and limitations of each approach is essential if policy-makers are to ensure that AI is used appropriately and effectively in health contexts.

Once a new technology is identified, there are several frameworks and conceptual models that can guide its introduction and adoption in different domains. Some of these emphasize technology-related questions, such as assessing the readiness of the technology and determining whether technical capabilities align with real-world applications. Others address field-specific considerations, such as evaluating the impact of an application on population health, patient outcomes or efficiency of health systems. Some emphasize feasibility and ethical concerns, which help answer value-based questions and identify barriers to adoption and implementation.

Within the health sector, health technology assessment (HTA) is "a systematic and multidisciplinary evaluation of the properties of health technologies and interventions covering both their direct and indirect consequences" (WHO, 2025b). This process includes summarizing and evaluating evidence on a range of considerations related to the use of health technologies, mainly relating to their implications for funding and delivering health care and for certain economic, social and ethical issues (WHO, 2025b). More details about the use of HTA in the context of AI are provided in section 3.1.2. Other examples include the diffusion of innovation theory (Rogers, 1962), the technology–organization–environment (TOE) framework (Tornatzky & Fleischer, 1990) and the more recently developed responsible AI frameworks (OECD.AI Policy Observatory, 2025).

A practical starting point when assessing the suitability of AI for a specific application or task is to consider three guiding questions that encompass population value, technology suitability, and broader appropriateness and feasibility considerations: 1) should this task be done? 2) can AI perform this

task? and 3) should this task be done by AI? This simple framework also aims to emphasize the role of AI as a tool to achieve a specific goal, rather than the goal in itself.

The first question pertains to field-specific considerations: should this task be done? In this context, this can be a health care, scientific or public health question. Does this application offer potential to improve population health? Does it align with the goals of the health system? Does it improve patient care?

There is now an extensive body of knowledge on the various ways to assess the impact of a policy, intervention or application, comparing the costs and benefits of different options, but an in-depth explanation of how to select or prioritize health interventions is beyond the scope of this book. This framework assumes that the task or application in question is ethical and consistent with international human rights standards.

The second is a technology-related question: can AI perform this task? As outlined earlier in this section, AI encompasses a range of models and approaches, each suited to particular types of tasks and uses.

A first step in answering this question will typically involve thinking through the nature of the tasks in question and the types of data available to address them. For example, identifying anomalies in chest X-rays involves classifying images as normal or abnormal, which can, in theory, be achieved by supervised machine learning models used with architectures that can process X-ray images. Typical machine learning tasks include classification, regression and clustering.

The second point to consider is the performance of the model and, specifically, how it compares with the current best practice. Often, the model performance can be evaluated in desk-based or in-silico settings, followed by real-world or prospective studies that assess clinical effectiveness and, ideally, long-term outcomes. Evaluation metrics should be selected carefully and reported following established guidelines to ensure transparency, reproducibility and relevance to clinical or public health outcomes (Jackson & Shortliffe, 2025). More detail on the evaluation of different types of AI models is provided in section 3.1.1.

Understanding the strengths and limitations of different AI techniques is essential to determine whether the task at hand aligns with what the technology can realistically and reliably achieve (Esteva et al., 2019; Janiesch et al., 2021; Taherdoost, 2023). However, as these technologies are complex and evolving rapidly, technology assessment bodies and regulators will have to recruit and retain teams with high levels of expertise in AI systems, supported by collaborations with relevant experts and institutions, while avoiding conflicts of interest.

The third question brings the other two questions together: should this task be done by AI? Answering this question involves a combination of value-based decisions and operational considerations.

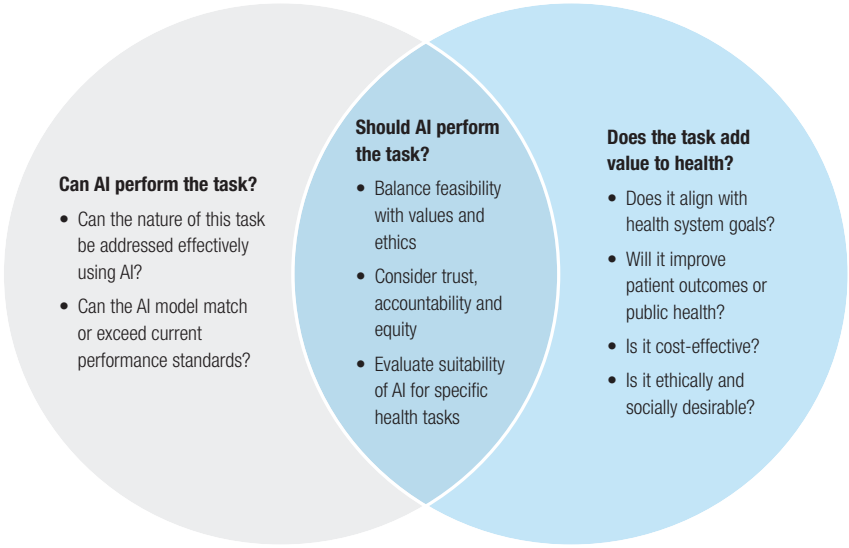
From an operational standpoint, it requires assessing the complexity and nature of the tasks, including their interdependence with elements of broader systems (inside and beyond the health sector), their frequency of occurrence and whether they rely on trust or human interaction. Repetitive tasks in areas that are data-rich are generally more amenable to automation, whereas specialized or relational tasks present greater challenges. Additional factors include data quality, information technology (IT) infrastructure and potential risks associated with task failure. High-stakes tasks, such as clinical decision-making, require much greater care, including consideration of ethical issues, than low-risk administrative functions, as reflected in the risk-based classification of AI systems in the EU AI Act. Ideally, most of these operational considerations should be assessed during the real-world evaluation of the AI technology as well as during the implementation and monitoring phases.

Several challenges to AI adoption will be explored in more detail throughout this chapter and in Chapter 3. They include limited evidence on the real-world performance of these models and their impact on long-term outcomes, as well as limitations in how performance metrics are reported. There are critical considerations regarding model bias and generalizability that can emerge across the AI life-cycle, including biases in the datasets used for training, which can perpetuate and reinforce existing societal and systemic discriminatory patterns. There are several challenges related to the adoption of these technologies and their integration into existing workflows, including the availability of infrastructure, data privacy considerations and staff capabilities. Additionally, issues around trust of the AI tools will be key, and include professionals and the public.

Broader concerns include the risk of exacerbating inequalities, environmental considerations, shifting power dynamics and other unintended consequences. Finally, regulatory considerations are fundamental to the development, evaluation and adoption of AI in health. While detailed legal analysis of specific applications falls outside the scope of this book, key regulatory considerations and frameworks are addressed in the introductory and concluding chapters.

In straightforward terms, these three questions can also be illustrated using a Venn diagram (Fig. 2.1, page 37). The questions in the left circle ask whether AI can perform the task. Those in the right circle ask whether the task being done by AI will add value for health. Those in the area of overlap ask, assuming the other answers are positive, whether AI should perform the task.

Fig. 2.1 *Applications of AI in health*



Source: Authors' compilation.

2.2 Overview of AI applications in health

Different ways have been proposed to classify and structure the scope of AI in health, such as by function (e.g. WHO public health functions) (Fisher & Rosella, 2022), by type of AI (e.g. types of predictive modelling applications) (Olawade et al., 2023) or by type of activity (e.g. knowledge generation) (Baclic et al., 2020). In this book, we classify the different applications across four broad domains: 1) health care delivery and patient care; 2) public health and health policy; 3) research and innovation; and 4) a cross-cutting theme of operational efficiency and resource management applications (Table 2.1).

This is not a comprehensive list of all potential applications, and some might span different health domains or build on each other, such as accelerated drug discovery and improving clinical trial processes. Moreover, these applications vary widely in their level of maturity, ranging from early-stage research to initial adoption and product implementation. Building on section 2.1, each application description in this chapter begins with an overview of the latest research and evidence, followed by a description of the relevant AI technologies involved, aiming to answer the question, can AI do this application or task? Looking at the operational and ethical considerations for each application, we will then be answering the question, should AI be used for this application or task? To avoid duplication, this discussion will centre on considerations most pertinent to each specific application, recognizing that several of these, such as model bias

and economic or environmental implications, are relevant to all AI models. In addition, given the breadth of the field and the rapid pace of advancement in AI in health, it is not possible to cover all potential examples or the emerging evidence. Despite these limitations, this framework will illustrate some of the most relevant and promising applications and serve as a starting point for decision-makers and professionals in the health sector.

Table 2.1 *AI in health applications*

Health care delivery and patient care	Public health and health policy	Research and innovation
Clinical decision support (diagnostic and risk assessment)	Understanding and forecasting population health	Accelerating drug discovery
Precision medicine	Behavioural insights	Improving clinical trial processes
Patient support tools	Infectious disease epidemiology	
Surgical robotics	Communications and engagement with the public	
Operational efficiency		
Automating routine tasks		
Resource allocation and decision-making		
Evidence synthesis		
Training and education of health care professionals		

Source: Authors' compilation.

AI offers the potential to reshape health care delivery and patient care by enhancing clinical decision-making, enabling more precise and personalized interventions and supporting both patients and professionals in a broad range of care settings. Its most prominent applications include clinical decision support, where AI models can aid in clinical diagnosis, especially in medical imaging, and in informing patient care by predicting the risk of specific health outcomes. AI can also play a central role in precision medicine, enabling care that is more tailored to the unique characteristics of patients or patient subgroups. By integrating and analysing different data, such as genetic, environmental and lifestyle variables, AI can help to develop customized treatment plans to improve patient outcomes and optimize health care delivery. Tools to encourage patient engagement have evolved from traditional rule-based systems to LLM-powered virtual assistants that can support self-management, chronic disease monitoring and mental health interventions. AI is also being explored as a means of supporting the training of health care staff, including using self-directed and adaptive learning platforms that can personalize content delivery, to enhance educational outcomes. Finally, in surgical practice, AI-integrated robotics have the potential to improve intraoperative precision, automating tasks and supporting real-time skill assessment, with applications extending across diverse specialities.

AI holds significant promise for advancing public health and health policy by enabling more integrated, data-driven approaches to understanding and improving population health. While many of its recognized benefits relate to efficiency, AI's potential in public health also lies in its ability to uncover previously unknown patterns and relationships that can inform more effective interventions. By expanding access to diverse data sources, such as electronic health records, social care systems, environmental monitoring and wearable technologies, and offering new analytical methods, AI can deepen our understanding of the determinants of health at both the population level and within specific groups, and enhance disease forecasting. A wide range of AI techniques can support population segmentation, risk stratification and the identification of complex, non-linear associations between exposures and outcomes, while also contributing to life-course research and advancing our understanding of multimorbidity. In infectious disease epidemiology, AI can enhance surveillance and improve disease modelling, with approaches such as graph neural networks offering insights into transmission dynamics. Additionally, AI can help apply behavioural science insights by analysing data from wearables and digital health tools to enable real-time monitoring and providing personalized advice aimed at supporting behavioural change and better understanding the impact of behaviour on health outcomes. Finally, AI can support communication and public engagement efforts through more targeted approaches, analysing population-level trends and facilitating tailored health messaging. Generative AI can also support multilingual, culturally appropriate communication and help counter misinformation. A recent review has examined how these various capabilities offer innovative ways to respond to a future pandemic (McKee et al., 2025b).

AI is emerging as a transformative force in health research and innovation, particularly within the drug development pipeline, where it offers significant potential to accelerate early-stage discovery and clinical trial processes. In the early phases of drug discovery, machine learning techniques are being harnessed to improve the identification of drug targets and to predict drug interactions and potential adverse effects, thereby streamlining the screening and selection of therapeutic compounds. These predictive capabilities are complemented by advances in generative AI, which enable the de novo design of novel molecules and protein structures through models that learn chemical and biological patterns in a manner analogous to language processing, ultimately expediting the generation of drug candidates tailored to specific therapeutic goals. Beyond discovery, AI also has the potential to reshape clinical trial processes by supporting trial design and execution through predictive modelling, virtual simulations using digital twins, and the generation of synthetic data to augment underrepresented cohorts or simulate control arms. Additionally, LLMs are being integrated to enhance data

synthesis, automate documentation and facilitate patient-trial matching, thereby reducing administrative burdens and enabling more targeted enrolment strategies.

Operational efficiency and resource management are critical domains in the health sector, with AI offering innovative solutions to streamline processes, reduce administrative burdens, optimize resource allocation and accelerate knowledge generation and synthesis. AI can automate administrative functions such as transcribing clinician–patient conversations, automating patient communication, scheduling appointments and processing billing information. In health care delivery settings, speech recognition and NLP algorithms can be particularly useful in documenting patient consultations and automatically generating outputs such as clinical summaries, structured documentation and standard letters. These advances have the potential to reduce the time spent on manual administrative duties, allowing health care professionals to focus on direct patient care and decision-making. At the organizational level, AI tools have the potential to enhance operational efficiency by forecasting changes in demand and supply, using both historical trends and real-time data. This enables hospital administrators to plan proactively and optimize resource allocation. In addition, machine learning techniques can deepen understanding of the underlying drivers of hospital workflows, supporting the design of targeted and evidence-informed operational improvements. AI can also facilitate priority-setting by supporting structured decision-making processes, including consensus-building platforms and by integrating novel data sources such as geospatial information. Finally, the ability of NLP algorithms and generative AI to extract, summarize, and generate insights from large amounts of text has the potential to revolutionize how researchers synthesize evidence to inform clinical and public health decisions.

2.3 Health care delivery and patient care

AI has the potential to reshape the landscape of health care delivery, offering new tools to support clinical decision-making, personalize care and enhance both patient and professional experiences. This section explores five key areas where AI holds significant promise: clinical decision support, where AI enhances diagnostic accuracy and predictive analytics; precision medicine, which leverages AI to tailor diagnosis and treatment to individual characteristics; patient support tools, including those that promote engagement, chronic disease management, and mental health support; training and education of health professionals, where generative AI is transforming learning environments; and surgical robotics, where AI is augmenting precision and intraoperative decision-making (Table 2.2, page 41). Together, these applications illustrate the breadth of AI's potential to improve outcomes, increase efficiency and support more responsive, patient-centred care.

Table 2.2 *Applications of AI in health care delivery and patient care*

Applications	Key findings	Considerations
Clinical decision support: enhancing diagnostic capabilities and predictive analytics	The application of AI models to medical imaging for diagnostic support and to medical records for individual-level prediction of outcomes or service use is one of the most mature and widely researched areas of AI in health. There is evidence that some of these models can outperform clinicians, and there are several diagnostic AI products in the market; however, most evidence focuses on early-stage development in controlled settings. Future directions in AI for clinical decision-making increasingly focus on integrating LLMs with electronic medical records to support clinical diagnosis and care, and using different data modalities as model input, such as text and images, through the use of LMMs	In most contexts, the use of AI in clinical decision-making will require human oversight and transparency, for example, through XAI. Adoption of machine learning and deep learning tools to support clinical decision-making in routine care has been historically slow until the recent emergence of generative AI tools. Reasons for this slow adoption include the complexity of clinical decision-making compared to task-specific AI models, data quality and availability, risks of biases in the models that reflect and perpetuate existing biases in health care systems and society, and complexities in human–machine collaboration. There are challenges when integrating some clinical decision-making tools into existing systems because they require real-time data streaming
Precision medicine: precision in diagnosis and personalization of treatment plans and care	While still in the early stages, AI holds promise for advancing the field of precision medicine by enabling more tailored, data-driven assessments that reflect individual variability, leveraging diverse and complex data sources such as clinical, genomic and behavioural data. Some of the most researched areas include AI-enabled adaptive therapy in precision oncology and AI-supported diabetes management, which illustrate how AI can optimize complex care pathways	Integrating AI into precision medicine requires robust clinical validation, transparent governance and safeguards for data privacy, especially given the sensitivity and fragmentation of genomic, behavioural and social data across public and private institutions. The growing interest in AI in precision medicine raises foundational questions about how best to prioritize resources, balancing efforts that advance increasingly individualized care with those that support broader interventions to improve population health
Patient support tools: patient engagement and accessibility, supporting chronic disease management and mental health support	The use of AI in tools to support patients' decisions has evolved from narrowly focused, rule-based systems to increasingly sophisticated and adaptive technologies leveraging generative AI. AI-powered virtual assistants and chatbots offer a promising way to expand access to health information and reduce pressure on health systems, particularly in settings with workforce shortages. Promising applications include chronic disease self-management and mental health support, and while some rule-based systems are more established, tools leveraging generative AI are in the earlier stages of adoption	Trust, digital literacy and the nature of patient–provider relationships are key factors influencing the adoption of AI-driven patient support tools. While LLM-enabled patient support tools enable more conversational and flexible interactions, they also pose significant risks related to misinformation, clinical accuracy and patient safety. Bias remains a pressing challenge, as AI systems trained on non-representative data or only one language may produce inconsistent outputs and deepen disparities in care. While AI may help expand access to health information in low-resource settings, its implementation also risks deepening disparities by limiting access to human health care professionals to those with greater financial resources

> > *continues*

Applications	Key findings	Considerations
Training and education of health professionals	Generative AI is increasingly being explored in the training and education of health professionals, building on earlier machine learning approaches focused on performance prediction and evaluation. LLMs are being used in medical education to support self-directed learning, including summarizing content, simulating clinical scenarios and clarifying concepts. Other AI tools are personalizing learning through adaptive platforms and virtual simulations, though most evidence is still early-stage and limited in scale	Evidence on the long-term impact of AI in health education is limited, with calls for more rigorous evaluation methods and clearer reporting frameworks. There are risks of overreliance on AI tools, which may affect critical thinking, clinical reasoning and increase exposure to biased or inaccurate outputs. AI use raises ethical and academic integrity concerns, prompting institutions to reconsider assessment practices and develop clearer guidance on acceptable use
Surgical robotics	AI has the potential to enhance the capabilities of surgical robotics by leveraging multimodal data from sensors, kinematic input and intraoperative imaging to support clinical decision-making and procedural accuracy. Applications include improving intraoperative visibility, automating surgical steps, tracking instruments and providing real-time evaluation of surgical skill, with most systems currently operating under human oversight	AI-enhanced surgical robotics face significant feasibility challenges, including limited access to high-quality annotated data, high infrastructure demands and the need for specialized skills and training Ethical concerns include safety, accountability and transparency, particularly in semi-autonomous systems where responsibility for errors may be unclear and model outputs are difficult to interpret Ensuring equitable and trustworthy use requires diverse training data, transparent consent processes and strong governance frameworks, especially to avoid reinforcing biases and to support safe clinical integration

AI: artificial intelligence; LLM: large language model; LMM: large multimodal model; XAI: explainable AI.

2.3.1 Clinical decision support

Enhancing diagnostic capabilities

One of the applications of AI that has attracted the most attention is its use in clinical decision-making, supporting clinicians when making diagnoses, or even, according to some enthusiasts, replacing them (McKee & Correia, 2025). This reflects the ability of AI to analyse vast amounts of clinical data quickly, accurately and consistently to identify patterns in complex datasets, such as medical images, genetic information and electronic health records, that may be too subtle or intricate for humans to detect easily. Advocates of greater use of AI argue that it can also help to reduce diagnostic errors by providing second opinions or flagging potential misdiagnoses, although, as we shall describe later, this is not as straightforward as it might at first seem. A particular attraction is its potential to expand access to diagnostic capabilities where specialists are scarce. This section explores the use of AI in diagnostics by examining the underlying technologies, their benefits and the current state of development.

AI techniques can support clinical diagnosis across a range of data types, with significant progress made in recent decades. One of the most advanced applications is in medical imaging, where AI can be used to detect, classify or localize abnormalities in images such as X-rays, MRIs and computed tomography (CT) scans. This progress is closely tied to advances in deep learning, particularly in computer vision. Models based on CNNs can be trained on large datasets of labelled images to automatically learn relevant features, achieving high accuracy in interpreting medical images. These models are tailored to the complexity and diversity of imaging across medical specialities.

A recent review of 30 studies of the use of AI in diagnostic imaging identified four areas where it could make a positive contribution (Khalifa & Albadawy, 2024). First, it could enhance image analysis, detecting minor anomalies and reducing human error while mitigating the effect of operator fatigue. For example, an interactive AI support system that enabled radiologists to click on parts of images that concerned them to generate a local cancer likelihood score improved their detection of breast cancer without increasing the time taken (Rodriguez-Ruiz et al., 2019). Second, it can reduce costs by improving efficiency and speed. Third, especially when using LMMs, it can combine different types of data on the patient in question to improve predictive and personalization capabilities. For example, a deep learning model was developed and validated to predict long-term incident lung cancer, combining data from chest X-rays, age, sex and smoking status (Lu et al., 2020). The model performance was superior to the United States standard, missing a third fewer lung cancers. Finally, through integration with existing data systems, such as electronic health records, it can enhance clinical decision-making by providing new insights and improving clinical workflow efficiencies.

Despite these advances in the use of AI in diagnostic imaging, challenges remain, such as the limited availability of high-quality labelled data. To address this, techniques such as transfer learning are used: models are first trained on large, general image datasets and then fine-tuned for specific medical tasks. Additionally, unsupervised and self-supervised learning methods have been developed to make use of unlabelled data, expanding the potential for training robust models (Esteva et al., 2021; Esteva et al., 2017). Another rapidly growing area is the use of generative AI models to perform tasks such as generating synthetic images to augment training data and enhancing image quality to improve diagnostic performance (Ibrahim et al., 2025).

Another important consideration is the so-called "black box" nature of many deep learning AI models, which can make it difficult to understand how decisions are made. This has led to growing interest in XAI, systems that provide insights into their reasoning processes. Studies have examined whether such transparency can

improve clinicians' confidence in AI outputs and enhance diagnostic accuracy. For example, one XAI system that analysed images of suspected melanomas provided explanations aligned with dermatologists' reasoning. This increased clinicians' confidence in the tool's predictions, although it did not improve diagnostic accuracy compared to the AI system operating alone (Chanda et al., 2024).

There is an extensive body of research evaluating the diagnostic performance of AI models compared to clinical standards across various imaging modalities and specialities, including cardiology, radiology and dermatology. A systematic review and meta-analysis found that the performance of deep learning models and health care professionals in detecting diseases in medical images across a range of specialities was comparable (Liu et al., 2019). Another systematic review and meta-analysis comparing AI and clinician performance with skin cancer diagnostics also found AI and expert dermatologist performance to be clinically comparable (Salinas et al., 2024). In another example, AI reading of mammograms for breast cancer screening was associated with a reduction in false positives and false negatives, with overall higher accuracy than radiologists alone (McKinney et al., 2020). A Swedish trial in which the second radiologist reading screening mammograms was replaced with an AI tool achieved a 4% higher cancer detection rate than the traditional double reading by radiologists (Dembrower et al., 2023).

Beyond medical imaging, advances in generative AI in natural language modalities are also shaping the way health professionals access information to inform diagnostic decisions. LLMs and LLM-powered AI agents can process clinical data, such as symptoms, vital signs and medical history, to suggest potential diagnoses and care options for health care professionals. (Zhou et al., 2025). A randomized controlled trial using clinical vignettes found no significant impact when LLMs were used to assist clinicians in their diagnostic reasoning. However, the LLM operating alone scored higher than either clinician group. (Goh et al., 2024).

If generative AI is to be used to support diagnostic reasoning, it must provide responses that are factually accurate, clinically relevant and consistent with established guidelines. One suggestion to facilitate this involves the use of techniques such as RAG. As noted in Box 1.4, this approach enables LLMs to draw on external, specialist knowledge in real time, incorporating locally relevant and up-to-date information while reducing the risk of erroneous or hallucinated content (IBM, 2023b). It should be noted that while there is now considerable evidence that LLMs can perform well in experimental situations, publicly available data from real-world use remains limited. Methods for evaluating LLM performance in clinical question-answering tasks are discussed in section 3.1.1.

Recent research has focused on the development of vision-language foundation models. These can combine visual (image or video) and textual (language)

information, allowing them to understand and generate both types of content together. By learning to associate visual elements with corresponding language descriptions, tasks such as image captioning, visual question answering, image-based reasoning and multimodal content generation, can be undertaken (Lu et al., 2024b). These can use structured (e.g. lab results, prescriptions) or unstructured (e.g. case notes) data from electronic health records (Wornow et al., 2023; Zhou et al., 2024a). Importantly, recent studies are also comparing the performance of CNN-based deep learning models with LLMs with multimodal capabilities, such as ChatGPT, in medical imaging analysis (Ahmed et al., 2025). As the AI field rapidly evolves, there might be a trend towards multimodal workflows and hybrid approaches between different types of models.

A key consideration in the use of AI for diagnostics is how it integrates with existing clinical workflows, along with broader implementation challenges. While much of the current literature focuses on the early stages of AI development, typically in controlled research settings, there is a growing body of work examining how these tools are being adopted in routine clinical practice. One notable milestone came in 2018, when the United States Food and Drug Administration approved the first autonomous AI diagnostic system for medical use. This system was designed to detect diabetic retinopathy without the need for clinician input, marking a significant step towards real-world deployment of AI in health care (Abramoff et al., 2018). Separately, the United Kingdom's National Institute for Health and Care Excellence (NICE) has reviewed several technologies that use varying types of AI to detect melanoma (NICE, 2022).

Predictive analytics: anticipating individual health risks and outcomes

Predictive analytic models can be used to inform patient care by estimating the likelihood of specific conditions or health events, both in and out of clinical settings. This can enable health care providers to intervene proactively, aiming to prevent disease progression and reduce health care costs. We will next briefly describe the existing and emerging applications of AI in predictive analytics to support clinical decision-making in health care, highlighting how machine learning and deep learning models are being used to enhance clinical risk prediction across a range of use cases.

Historically, risk prediction tools for clinical decision-making were developed using data from large cohort studies, for example, to help predict the individual risk of cardiovascular disease (The Framingham Heart Study, 2025). More recent advances in risk prediction modelling use real-world data from electronic health records. This has the advantage of including a wider range of individuals than the often select group of those recruited to and retained in cohort studies, while the much larger numbers involved enable prediction of a wider range of clinical

outcomes using computational techniques such as generalized linear models and machine learning models (Goldstein et al., 2017). Other examples include using AI models to predict the risk of readmission of patients with multimorbidity (Guo et al., 2023; Hassan et al., 2024) or with heart failure (Preiksaitis et al., 2024), to estimate the length of stay for people with learning disabilities and multiple long-term conditions (Abakasanga et al., 2025) or to predict the likelihood of developing type 2 diabetes by analysing electrocardiograms (Pastika et al., 2024). Additionally, AI models have been developed to help predict survival outcomes, such as for cervical cancer patients, based on pathological images (Zhang et al., 2023), or hepatocellular carcinoma, based on patient records (Li et al., 2023b; Wang et al., 2021).

These benefits are greatest where electronic records can be linked to follow patients over time, enabling the training of models. In more recent years, longitudinal patient data have been used to train deep learning models that can predict patient health outcomes from information collected earlier in the patient's journey. For instance, Google DeepMind trained a deep learning model on over 700 000 records to predict inpatient episodes of acute kidney injury and to compile a list of the clinical features that are most relevant to each prediction, and projected future trajectories associated with results of certain blood tests (Tomasev et al., 2019). The ability of LMMs to combine data from different sources, such as image and text, offers many new opportunities, as with a study describing the development of a multimodal deep learning prognostic model using whole-slide images and tumour stage as an input to predict distant recurrence of endometrial cancer (Volinsky-Fremond et al., 2024).

NLP and LLM applications have also been described in this field to leverage unstructured data from clinical notes as an input to the prediction models. This approach was examined in early sepsis prediction and diagnosis by combining structured data from the electronic health record and unstructured clinical notes, extracted using an NLP-enabled AI algorithm. This approach enabled an increase in early detection of sepsis and a reduction in false positives compared to clinician predictions, and the incorporation of unstructured text improved the accuracy compared to using only structured data for early warning of sepsis onset (Goh et al., 2021). Another study reported significant improvements in the ability to predict perioperative complications by taking text from clinical records (Alba et al., 2025). The authors fine-tuned LLMs using perioperative clinical notes, further improving accuracy. Incorporating outcome labels through further fine-tuning added another layer of refinement. The greatest improvement was with a foundation model trained using multi-task learning across six postoperative risks, which outperformed all previous approaches. Finally, integrating structured tabular data, such as lab results and patient history, improved predictions further, especially for rare complications like pulmonary embolism. This layered approach

demonstrated that adapting LLMs to domain-specific language and combining them with structured data can significantly improve early risk detection in surgical care.

Should AI be used to support clinical decision-making?

AI tools are particularly well-suited to streamline repetitive, well-defined tasks within clinical workflows, especially in areas such as clinical decision-making. Some of the most advanced and mature applications of AI in health are in clinical decision support, ranging from image-based diagnostic tools to models that use health records to predict outcomes or service use. However, while many of these techniques, including shallow machine learning for risk prediction and deep learning for image classification, have been in development for years, their adoption in health systems has been slow. In addition, there tends to be a gap between their reported performance during the early research stage and their actual performance in real-world implementation, and evidence on their long-term impact in real-world settings remains limited. The following section outlines key factors that help explain some of the challenges in scaling and integrating these technologies into routine care.

System readiness and integration considerations

A fundamental question for those considering the adoption of AI to enhance clinical decision-making is whether it can improve on the performance achieved by current clinical standards, but the answer is not straightforward. On the one hand, despite the growing number of studies comparing AI tools to health professionals, assessing the performance of AI tools in a way that reflects the complexity of real-life clinical decision-making is extremely challenging. Evaluation of different types of AI tools and their respective performance metrics will be covered in more detail in section 3.1.1. On the other hand, there is a growing body of research on how the performance of AI tools, when used by health professionals, can vary, influenced by factors such as their level of clinical experience and their trust in the AI tools. In some studies, for instance, less experienced clinicians seemed to benefit more from AI tools compared to those with more experience (Chanda et al., 2024; Reverberi et al., 2022). However, a recent study examining the effects of AI assistance on radiologists across chest X-ray diagnostic tasks highlighted the complexity in understanding the specific factors that influence individual diagnostic performance in human–AI collaboration. For example, it found that years of experience and familiarity with AI tools failed to predict the impact of AI assistance, and that AI errors strongly influenced radiologist performance (Yu et al., 2024). Similarly, other studies emphasize the importance of assessing not only overall accuracy, but also the different types of errors that can emerge in AI–human collaboration in medical decision-making (Rosenbacke et al., 2024b).

A second critical issue in developing models that guide clinical decision-making is data quality and model drift. Predictive models that rely on electronic health records and patient history suffer from a crucial limitation in that these records are often incomplete or may have errors. Moreover, the data on which they rely is typically episodic rather than continuous over time (i.e. data from most lab tests is only at specific time points). These limitations can only be addressed by long-term investment in the underlying health care data infrastructure to ensure high-quality data for training or fine-tuning AI models. The alternative is reliance on models that have been trained with data from other settings, which might not be generalizable to the target population.

Another key operational consideration is how AI models are integrated into existing workflows, data infrastructure and decision-making processes. Importantly, most machine learning models are trained to perform a single, narrowly defined task, for example, predicting a specific outcome or classifying an image according to the presence or absence of a particular feature. In contrast, clinical decision-making is inherently more complex and holistic, requiring professionals to evaluate a broad range of diagnostic possibilities when interpreting information. While this limitation is comparable to that of many existing diagnostic tests, such as point-of-care tests that detect a single pathogen or biomarker, it nonetheless constrains the extent to which AI tools can be seamlessly incorporated and scaled in routine clinical care.

Governance, ethics and rights considerations

One of the main concerns in the use of AI tools in clinical decision-making is the risk of bias, which can arise at several stages of the AI life-cycle. One of the key sources of bias is the training data used to develop AI models, which can be incomplete or unrepresentative of the population to which it is being applied. On the one hand, this can reduce the performance of the AI model. On the other hand, there is a growing body of evidence showing how biases stemming from AI model risk can perpetuate and amplify the many existing biases that are rooted in society and medical institutions, disadvantaging and discriminating against some communities (Stevens & Keyes, 2021).

A recent study showed how one of the most widely used AI models for reading chest X-rays consistently underdiagnosed lesions in certain individuals (e.g. younger, female and black, and especially in those combining these characteristics) over a wide range of pathologies and datasets, and to a greater extent than when readings were by radiologist (Yang et al., 2025b). When using AI models to support imaging diagnostic decisions, skin colour bias is a well-known limitation (Bencevic et al., 2024). The performance of AI models on darker skin tones is generally inferior to that on lighter skin tones, reflecting

limited research in this area and the underrepresentation of images with darker skin tones in the training datasets, particularly at the scale required to develop deep learning and foundation models (Zhou et al., 2024a).

Another important consideration when implementing AI tools is to understand the stakeholder perspectives of those involved or impacted by the technology. A systematic review of qualitative studies of the implementation of diagnostic AI argued for meaningful engagement with patients, clinicians, researchers and health care leaders (Ling Kuo et al., 2024). The studies reviewed found that stakeholders held diverse yet interconnected views on the adoption of diagnostic AI. Patients expressed a strong desire to retain the human touch in care, valuing empathy, eye contact and the ability to ask questions, all qualities that they felt AI lacks. Their acceptance hinges on trust, data privacy and maintaining autonomy over health care decisions. Clinicians, meanwhile, were cautiously optimistic but emphasized the need for control, explainability and integration into existing workflows. They worried about liability, deskilling and the potential erosion of their professional roles. Researchers focused on developing AI that clinicians would actually use, stressing the importance of high-quality, representative data and collaboration with medical professionals. Health care leaders viewed AI as a solution to workforce shortages and inefficiencies but recognized that successful implementation depends on clinician engagement and robust infrastructure. Across all groups, trust, transparency and collaboration were consistently identified as essential for successful adoption. The authors proposed a modification of the NASSS framework to guide implementation of diagnostic AI; NASSS stands for non-adoption, abandonment, scale-up, spread and sustainability (Greenhalgh et al., 2017). This identifies a series of challenges facing those adopting health care technologies across seven domains: the condition of interest, technology, value proposition, adopter system, organization, wider context and embedding/adaptation over time.

Another unintended consequence of introducing AI into clinical decision-making processes is the potential erosion of health care professionals' reasoning skills and foundational knowledge. As these tools become more widespread and professionals increasingly rely on AI to automate tasks and generate insights for patient care, there is a risk that these essential skills will diminish over time. Notably, the impact is likely to vary depending on the stage of a health care professional's career and their prior experience, with those in the earlier stages, who have not yet fully developed foundational knowledge, being potentially more vulnerable.

In summary, AI has the potential to reshape the health care landscape of diagnostics and predictive analytics, offering unprecedented opportunities for

improving patient care. By leveraging its strengths in data analysis and pattern recognition, AI can enable earlier detection of diseases and proactive health management. However, ethical and practical challenges remain and must be addressed if the benefits of AI are to be realized equitably and sustainably.

2.3.2 Precision medicine

Precision medicine and personalized medicine are terms that are often used interchangeably; however, precision medicine typically refers to the tailoring of medical treatment to specific subgroups of patients based on genetic, environmental and lifestyle factors, while personalized medicine encompasses a broader, more holistic approach that integrates not only biological variables but also psychosocial, behavioural and patient preference factors. This nuanced distinction reflects the differing emphases of the biomedical and biopsychosocial models of health (Delpierre & Lefevre, 2023). While AI holds promise for advancing both domains, its most measurable and immediate impact is currently observed in the field of precision medicine. In this area, AI can enable medical care to be tailored to the unique characteristics of patients or patient subgroups. By leveraging vast datasets and advanced computational techniques, AI models hold promise to combine genetic, environmental and lifestyle factors to design customized treatment plans, improving patient outcomes and optimizing health care delivery. Its strength lies in its ability to process and extract insights from the immense scale and complexity of biomedical data that traditional methods cannot manage efficiently.

Precision in diagnosis

As previously described, research and development of AI-powered diagnostic tools are rapidly evolving. Within precision medicine, AI can further enhance diagnostic capabilities by helping to uncover individualized patterns and stratify patients into biologically meaningful subgroups (Ahmed et al., 2020). Traditional diagnostic approaches often rely on aggregate population-level evidence and standardized clinical guidelines, which may not capture the biological and contextual variability between individuals, although these should usually be interpreted using clinical judgement and knowledge of the individual patient. AI solutions have the potential to identify subtle, individualized patterns in high-dimensional data sources, such as genomic sequences, radiological images, electronic health records and wearable sensor outputs, that might not otherwise be obvious (Johnson et al., 2021). These insights have the potential to enable not only more accurate diagnoses but also more nuanced predictions of disease subtype, prognosis and likely treatment response, which are core elements of precision medicine.

One of the most prominent applications of AI within precision medicine is in the field of oncology, driven partly by the toxicity and risk of side-effects from many cancer treatments, as well as the potential to target specific tumour molecular targets (Dugger et al., 2018). Precision oncology aims to enhance diagnostic accuracy and optimize treatment strategies by leveraging multidimensional datasets, including genomic, histopathological, radiological and clinical information, to gain deeper insights into tumour biology and molecular pathways. A review by Fountzilas et al. (2025) highlighted several key applications of AI in this domain. These include digital pathology, where AI is used to automate immunohistochemistry scoring and to extract novel insights from tumour tissue analytes; digital radiology, particularly in predicting responses to immunotherapy (using radiotracers to characterize the microenvironment of the tumour); and integrative multimodal analyses that combine diverse data types to improve prognostic classification and treatment stratification.

Personalization of treatment plans and care

AI can enhance personalized medicine by optimizing treatment plans based on individual patient data. Machine learning models can help predict how patients respond to treatments, accounting for genetic predispositions, medical history and lifestyle factors. This has the potential to enable clinicians to adjust dosages, select alternative therapies or combine treatments to achieve optimal outcomes.

In cancer treatment, adaptive therapy strategies have been proposed as a way to control tumour growth and delay drug resistance by dynamically adjusting therapy based on the tumour's response over time. This method leverages evolutionary principles, seeking to maintain a stable population of drug-sensitive cells to suppress resistant ones, thereby delaying or preventing treatment failure. Unlike traditional maximum-tolerated dose strategies, adaptive therapy uses dynamic dosing or intermittent treatment to reduce selective pressure and limit resistance. In this context, deep reinforcement learning techniques have been proposed to deliver dynamic, patient-specific treatment scheduling. In a study investigating this approach with prostate cancer patients, deep reinforcement learning was found to increase the time to relapse compared to the usual standard of care (Gallagher et al., 2024).

Similarly, in diabetes management, AI-driven platforms can analyse glucose levels, dietary habits and activity patterns to provide personalized recommendations for insulin dosing and lifestyle adjustments. For example, machine learning and deep learning models have been used to analyse electronic health records and predict treatment failure, and reinforcement learning approaches have been proposed to optimize insulin regimens (Sheng et al., 2024; Tarumi et al., 2021; Wang et al., 2023b). AI tools have also been proposed to support health care professionals in

delivering tailored diabetes patient care, leveraging LLMs and electronic health records to respond to patients' needs (Kim et al., 2025). In cardiology, research is ongoing to better understand how AI can provide personalized risk assessment and care plans for those people living with chronic heart failure by tailoring care models to each patient's risk profile with multi-source cardiovascular data (AI4HF, 2025).

Should AI be used in precision medicine?

AI has the potential to transform precision medicine by uncovering hidden patterns in complex datasets, enabling better classification of patient subgroups and informing more targeted treatment decisions. While much of the progress in the past decade has been driven by advances in AI-based image processing, recent developments in LLMs and multimodal AI are expected to accelerate the evolution of precision medicine further (Truhn et al., 2024). However, these emerging technologies require more robust clinical evaluations, and their adoption raises important ethical, regulatory and implementation challenges.

System readiness and integration considerations

Both AI and precision medicine remain rapidly developing fields. Despite substantial investment and scientific enthusiasm, many of their expected benefits, such as improved clinical outcomes and health system efficiency, have yet to be consistently demonstrated in real-world practice. To responsibly integrate AI into precision medicine, regulatory frameworks and governance systems must continue to evolve. This includes establishing standards for validating AI models in clinical contexts and ensuring they can be safely monitored and updated as new data emerges. Regulators and professional bodies are stepping up efforts to address these needs, but comprehensive and practical guidance remains limited. Importantly, cross-disciplinary collaboration among clinicians, data scientists, ethicists and policy-makers will be essential to ensure that AI tools are not only technically sound but also aligned with principles of fairness and accessibility.

Governance, ethics and rights considerations

At the core, these innovations raise critical ethical questions about the purpose of health care systems, considerations that extend beyond specific AI questions and touch on foundational questions about the role of precision medicine more broadly: should finite resources be directed towards increasingly individualized care, or towards interventions that improve health outcomes of the entire population? The vision of tailoring care to an individual's genetic, behavioural and environmental profile is compelling, yet the infrastructure, data and algorithmic development required to achieve this at scale demands substantial investment, with accompanying opportunity costs (Litvinova et al., 2025). This prompts important questions about health equity and societal benefit: if AI enables

unprecedented personalization of care, will only those in well-resourced health systems or with access to digital tools benefit? Without deliberate strategies to ensure inclusivity, there is a real risk that precision medicine could deepen health disparities rather than reduce them.

As discussed throughout this book, AI models are inherently shaped by the data on which they are trained. In the context of precision medicine, this can include genomic, imaging and behavioural data sources that may be fragmented, incomplete or biased. If these data lack diversity or reflect systemic inequities, the resulting models risk producing inaccurate or harmful recommendations. Transparency and explainability are also critical, particularly when AI is used to guide high-stakes decisions about personalized treatment.

In addition to ethical and resource allocation concerns, there are significant data governance challenges to overcome before AI can realize its full potential in precision medicine. A major concern is data privacy: developing AI tools that enable personalized care requires the integration of highly sensitive information, including genomic, behavioural and social data. Without stringent safeguards, this raises the risk of data breaches or misuse. These risks are compounded by the fragmentation of data sources across public and private institutions, making secure and coordinated data access and governance even more complex.

2.3.3 Patient support tools

The use of AI in tools to support patients' decisions has evolved from narrowly focused, rule-based systems to increasingly sophisticated and adaptive technologies leveraging generative AI. Early applications largely relied on expert systems and decision trees, and leveraged NLP to offer structured support for tasks such as symptom triage, medication reminders and answering frequently asked questions (Laranjo et al., 2018; Laymouna et al., 2024). While these tools have demonstrated utility in low-complexity scenarios, their capacity to engage with patients meaningfully or adapt to varied contexts was limited (Tudor Car et al., 2020). Recent advances, particularly the emergence of LLMs, have expanded the potential of AI in this area, enabling more nuanced, conversational and context-aware interactions (Ayers et al., 2023). Applications currently extend into areas such as chronic disease self-management, mental health support and integration with wearable technologies (Clark & Bailey, 2024). However, while early evidence suggests promise in enhancing patient engagement and accessibility, most LLM-based tools remain in experimental or pilot phases, with ongoing concerns about safety, accountability and clinical integration (Huo et al., 2025).

Enhancing patient engagement and accessibility

Virtual assistants and chatbots have the potential to expand access to health care information while reducing pressure on health care providers and systems. By offering 24/7 availability and immediate responses to health-related queries, these tools can serve as a first point of contact for individuals seeking guidance. In the context of widespread health care workforce shortages and the growing tendency for individuals to consult the Internet for health advice, well-designed and clinically validated symptom checker tools present a particularly promising solution. Their utility is especially pronounced in settings where access to in-person care is constrained but where the necessary digital infrastructure is in place to support remote health technologies. However, some caution is needed as many of these products have been developed to work primarily in English and are likely to perform less well in other languages. There are some specific challenges with certain European languages (Box 2.1, page 55). If these challenges are to be addressed effectively, long-term investment to create high-quality corpora for underrepresented languages may be required, including the inclusion of local dialects and code-switching patterns to improve natural language understanding (Uebler, 2001). Tools such as Meta's No Language Left Behind project show promise, but remain unreliable for sensitive health care communication (Adelani, 2024).

A recent rapid review assessing the role of chatbots in health care identified a broad range of applications supporting patient care and self-management (Laymouna et al., 2024). These included mental health support, counselling, provision of treatment advice and tools for self-management and monitoring of chronic diseases. A smaller subset of applications focused on functions such as triage, screening, risk assessment, referral and appointment reminders. The review also highlighted the potential for chatbots to promote health literacy and encourage healthy behaviour change. Notably, the majority of the tools reviewed were AI-based, reflecting a growing trend towards the use of conversational interfaces in digital health interventions.

Box 2.1 *Linguistic challenges when using voice recognition and chatbots*

Estonian, Finnish and Hungarian: These languages have an extremely rich morphology (a single verb can have hundreds of forms), long compound words and vowel harmony so systems need much more data to cover inflected forms.

Polish: This language has dense consonant clusters, nasal vowels and frequent homophones. For example, words like *wstrząs* (shock) pack in many consonants with subtle timing cues.

Irish (Gaelic): In this language there are large differences between spelling and pronunciation, plus initial consonant mutations that change the start of words depending on grammar. For example, *bh* can sound like /v/, /w/ or be silent, depending on context.

Danish: There is a large gap between spelling and actual speech; many unstressed syllables are reduced or dropped. The "stød" (a glottalization feature) can distinguish words but is subtle acoustically.

Swiss German dialects: Dialects differ greatly from Standard German, with shifting vowel systems, local vocabulary and no standardized spelling.

Icelandic: As there is only a small population speaking this language, there is less training data. The language also preserves archaic sounds and complex inflections.

Scottish Gaelic and Welsh: These languages have some of the same issues as Irish, while mutation systems and small corpora make accurate modelling difficult.

Basque: This language has unique phonotactics, an agglutinative structure and relatively few large public datasets.

Maltese: Most early Maltese chatbots were rule-based and struggled with natural conversation and code switching (mixing Maltese and English in the same sentence). There are also limited training data.

Romani: Romani has many dialects with significant differences in vocabulary and grammar, while language barriers are compounded by historical mistrust and a lack of culturally appropriate messaging.

Sámi languages (e.g. Northern, Lule, Southern Sámi): These languages have distinct orthographies and are spoken by relatively small populations, and they are spoken by few health professionals, meaning that many technical words are untranslatable.

Over the past decade, a range of patient-facing tools, primarily in the form of mobile applications, have been developed and introduced to the market. These tools offer functionalities such as symptom checking, preliminary medical advice, triage and administrative support, with their scope often determined by the degree of integration with existing clinical pathways. While some of these applications have been incorporated into public health systems, concerns have been raised regarding the rigour with which their efficacy, safety, cost-effectiveness and clinical impact are evaluated (Fraser et al., 2018). A comparative study assessed the diagnostic accuracy of general practitioners versus eight AI-powered symptom assessment applications (Gilbert et al., 2020). The study found that while some approached general practitioner-level performance in advice on urgency, none matched general practitioners in diagnostic accuracy. Coverage varied widely, with some excluding users based on age, pregnancy or symptom severity, limiting their usefulness. Diagnostic accuracy was inconsistent, with the top-3 condition suggestion accuracy ranging from 23.5% to 70.5%, compared to 82.1% for general practitioners. Applications that excelled in one area often underperformed in others. This study highlighted the need for real-world testing and standardized evaluation, as current performance is uneven and sometimes unreliable.

Despite these challenges, LLM-powered public-facing chatbots have the potential to replace general search engines as primary tools for seeking health information, particularly given their capacity to operate in multiple languages. Similarly, these tools may increasingly play a central role in self-diagnosis and preliminary health assessment. The evolving capabilities of LMMs could enable the integration and interpretation of diverse data sources, offering more personalized and comprehensive support (WHO, 2021). Frameworks are being proposed to benchmark the capabilities of LLMs for health advice across different criteria (OpenAI, 2025b).

Supporting chronic disease management

The global burden of chronic diseases such as diabetes, hypertension, cardiovascular disease and chronic respiratory conditions continues to rise, placing immense pressure on health care systems. Managing these conditions requires sustained patient engagement, self-monitoring, adherence to long-term treatment regimens and behavioural modifications, all of which can be difficult to maintain without ongoing support. Patients in many settings face challenges such as limited access to regular follow-up and fragmented care. AI-powered patient-facing tools, including chatbots and virtual assistants, may help to bridge these gaps. These tools can support behaviour change, deliver timely health information and facilitate remote monitoring.

A scoping review exploring the use of AI tools for self-management of chronic conditions identified that most applications focus on medical and behavioural self-management, and just a few on emotional self-management. Conversational AI was the most widely used technology, but most tools remained in development or an early testing phase (Hwang et al., 2025). A systematic review and meta-analysis of examples in diabetes assessed the use of chatbots to support self-management and found that most focused on educational and management support on diet, exercise, glucose monitoring, treatment and risk of complications. Their use was often effective in reducing blood glucose, but not in reducing weight. However, the quality of the evidence was low (Wu et al., 2024). AI-powered chatbots can also be integrated with wearable devices, enabling continuous monitoring of vital signs and providing health coaching for chronic conditions, such as diabetes and cardiovascular disease.

Mental health support

The global mental health landscape faces significant challenges, with severe shortages of mental health professionals, especially in low- and middle-income countries, exacerbated by barriers such as stigma and limited access to care. AI-powered chatbots offer a potentially scalable solution to bridge these gaps by providing immediate and anonymous support. These tools can deliver interventions grounded in evidence-based therapies, such as cognitive behavioural therapy, and are accessible via smartphones and other digital platforms, making them particularly valuable for individuals hesitant or unable to seek traditional in-person therapy (Abd-Alrazaq et al., 2019; Haque & Rubya, 2023). While these tools are not considered replacements for licensed therapists, they can play a role in monitoring symptoms, offering low-intensity support and potentially guiding users towards professional care when needed.

A systematic review and meta-analysis of AI-based conversational agents designed to promote mental health and well-being found that these tools significantly reduced symptoms of depression and psychological distress. However, the interventions did not lead to significant improvements in overall psychological well-being. Most of the tools examined were retrieval-based, where responses are selected from a pre-established repository, rather than generative AI. In addition, their primary functions centred on delivering psychotherapy and psychoeducation, followed by offering social companionship or practical assistance. Depression and anxiety were the most commonly assessed outcomes across the studies (Li et al., 2023a). In England, NHS mental health services have adopted Limbic, an AI-powered conversational chatbot that supports assessments and referrals in psychological therapies, screening, signposting and asking clinical questions.

An observational study of this tool found that services utilizing it had increased referrals, and this effect was more pronounced among minorities. Potential drivers of improvement included the lack of human interaction and patient's self-realization of treatment need (Habicht et al., 2024; NHS Transformation Directorate, 2022).

The technologies underlying these tools have evolved rapidly. As previously described, early chatbots relied on rule-based logic and scripted dialogue trees, which limited their responsiveness and adaptability. Subsequent advances in NLP enabled more fluid conversations and expanded the chatbot's capacity to recognize diverse forms of input. More recently, LLMs and generative AI have further enhanced chatbot capabilities, allowing for context-aware and personalized interactions. However, the incorporation of LLMs into mental health applications has been approached cautiously due to concerns about safety, misinformation and ethical risks. Some therapy chatbot companies, for example, have described the need to rely on rule-based systems for core therapeutic interactions but have also been exploring hybrid models that selectively integrate generative AI while maintaining tight controls to prevent hallucinated or harmful responses (IEEE, 2024).

Should AI be used in patient-facing support tools?

The growing use of AI-powered virtual health assistants and chatbots in patient-facing roles presents significant ethical and feasibility challenges. While these tools offer promising opportunities to expand access, support self-management and reduce strain on health systems, their responsible implementation requires careful attention. Key considerations include the accuracy of the information provided, equity of access, bias in the models, digital literacy and the impact on trust and interpersonal relationships.

System readiness and integration considerations

A key concern is the accuracy and clinical reliability of the information provided. There is considerable variability in the diagnostic and triage performance of symptom checkers, and there are weaknesses in the design and implementation of many evaluations. For instance, as explained in more detail in section 3.1.1, evaluations of LLM-based applications often focus on their ability to answer standardized medical questions rather than their effectiveness in real-world clinical contexts using patient data. While the conversational fluency of LLMs offers opportunities to create more natural, human-like interactions that may improve user engagement and trust, it also brings new risks. Fundamentally, there is potential for misinformation or hallucinated responses, which are outputs that appear confident and plausible but are factually incorrect. Without consistent

validation in diverse clinical settings, there is a danger that such outputs could lead to inappropriate decisions, adverse patient outcomes and erosion of trust in digital health tools. A recent benchmark of LLMs for health advice conducted by OpenAI identified that models had severe limitations in seeking the necessary context to provide precise responses (OpenAI, 2025b).

Other concerns about feasibility relate to trust, digital literacy and the nature of patient–provider relationships. Trust remains a fundamental component of effective health care, and many patients may prefer human interaction, particularly when dealing with emotionally sensitive or complex issues. The therapeutic relationship plays a vital role in engagement and adherence, a function that automated tools cannot fully replicate. Additionally, a wide variation in baseline digital and health literacy may limit the accessibility and impact of chatbot-based interventions. Individuals who struggle with digital tools may misinterpret information or struggle to use applications effectively, leading to confusion or reduced benefits. Ultimately, full integration into health care systems necessitates alignment with clinical workflows, adherence to relevant regulations and thorough evaluation of safety, cost-effectiveness and long-term implications. Without clear standards and governance, the widespread adoption of these tools may outpace our understanding of their consequences

Governance, ethics and rights considerations

Equity and bias are among the most pressing ethical concerns. On the one hand, AI tools can enhance access in underserved areas where no alternatives exist, particularly in regions with acute health workforce shortages. However, there is also a risk of deepening inequalities, whereby individuals in lower-resource settings receive care primarily through automated systems, while those in wealthier contexts continue to benefit from direct access to health professionals. The limitations of the training data further exacerbate these concerns. If AI systems are not trained or regularly updated with datasets that reflect the diversity of the populations they serve, they may produce biased or inconsistent outputs, potentially exacerbating disparities in care. Differences in language, cultural background or demographic factors could result in divergent responses to similar queries. Tools such as Limbic have demonstrated potential to improve access among historically underrepresented groups, but this requires careful, inclusive design and implementation.

In summary, virtual health assistants and chatbots have the potential to reshape health care delivery by enhancing accessibility and supporting the management of certain conditions. However, addressing challenges related to accuracy, privacy and patient trust is essential to fully realizing their benefits.

2.3.4 Training and education of health professionals

Several applications of AI are being explored to support the training and education of health care professionals across various domains. While early work focused on the use of machine learning approaches to predict and evaluate students' performance (Winkler-Schwartz et al., 2019), the rapid changes in access to generative AI tools are having a profound impact in the way knowledge is gained and tested. Several literature reviews have summarized the latest evidence on opportunities and risks of using AI tools in the training and education of health professionals (Gordon et al., 2024; Preiksaitis & Rose, 2023; Safranek et al., 2023). Key applications include the use of generative AI for self-directed learning and exam preparation, as well as adaptive learning platforms that can personalize content delivery.

Generative AI, especially LLMs, is being widely explored for its potential to assist with self-directed learning. These models can generate natural language responses to medical or other health care professionals' questions, provide explanations and simulate interactions, thereby offering the possibility to support and enhance independent study. Evidence from recent reviews shows that medical students and educators are already using LLMs in diverse ways, including for summarizing content, clarifying concepts, generating lists of differential diagnoses, simulating interactive clinical conversations and developing or reviewing simulated exam questions (Safranek et al., 2023).

In addition to LLMs, other types of models and adaptive learning platforms are being developed to optimize the learning pathway by identifying gaps and adjusting the material in real time. Some of these platforms are leveraging deep learning techniques to simulate virtual patients and support students with diagnostic reasoning (Furlan et al., 2021; Lin et al., 2025) and to tailor educational experiences to an individual's cognitive load (Ruberto et al., 2021). AI models have also been used to augment virtual reality and virtual patient simulation in the training of communication skills (Stamer et al., 2023). This is consistent with a shift in student learning towards AI-assisted engagement with course material. However, much of the current evidence is exploratory or based on small-scale implementations, with limited longitudinal data to confirm learning outcomes (Feigerlova et al., 2025).

Should AI be used in the training and education of health professionals?

While there is growing interest in the applications of AI and particularly generative AI to support the training and education of health professionals, there are several considerations and risks. These include limited evidence on the impact of these tools on learning outcomes, skillset erosion and risks of biases and hallucinations in the models.

System readiness and integration considerations

While enthusiasm for AI tools in training and education of health professionals is high, the evidence base remains limited in terms of rigorous evaluations and demonstrable impact on long-term learning outcomes. As noted by a systematic review focused on assessing the impact of AI on educational outcomes in health professionals' education, there are several areas for improvement in the way studies are designed and a need for further guidelines on how to evaluate the quality of this type of research (Feigerlova et al., 2025). Gordon et al. (2024) have highlighted the existing challenges in reporting AI innovations in medical education and have proposed a framework that focuses on form, use case, context, education form, technology and level of technological integration (the FACETS framework).

Governance, ethics and rights considerations

A key concern is the potential for erosion of clinical reasoning and foundational knowledge. As generative AI tools such as LLMs become more integrated into students' workflows, for example, by generating summaries, explanations or draft responses, there is a risk that students may become overly reliant on AI outputs rather than developing their critical thinking and decision-making abilities. This risk is especially pertinent in clinical education, where the ability to navigate uncertainty and synthesize diverse information is crucial. Automation bias can also be a problem, occurring when individuals assume that AI-generated content is accurate.

Another fundamental issue is the risk of biased or inaccurate outputs. Generative AI models are trained on large datasets that may reflect historical or systemic biases, particularly in the case of LLMs trained on general knowledge instead of context or field-specific topics. This can lead to outputs that perpetuate stereotypes or present clinically inappropriate advice. Moreover, hallucinations, where the model fabricates plausible but incorrect information, remain a fundamental issue of the generative AI problem, especially in models not specifically fine-tuned for clinical accuracy. AI-generated scenarios that lead to misinterpretation of essential health care topics can severely impact the quality of education and ultimately patient care.

Academic integrity can also be a significant challenge. The ease with which AI tools can generate high-quality text or simulate exam responses raises questions about plagiarism, authorship and the authenticity of learner assessment. Additional concerns include the possibility that students may use generative AI tools to bypass meaningful engagement with coursework, particularly in written assignments or reflective tasks. Institutions are now being prompted to reconsider their assessment strategies and develop guidance on acceptable AI use in education.

2.3.5 Surgical robotics

Robotics has transformed some areas of surgery by offering enhanced precision, flexibility and safety, combining the strengths of humans and machines (Ciuti et al., 2025). These advanced systems use robotic arms, high-definition cameras and sophisticated software to assist surgeons in performing complex surgeries with minimal invasiveness. Robotic platforms have also improved in ergonomics and haptic feedback, offering surgeons better control and reducing fatigue during long operations, although the impact of robotic surgery on cognitive workload may require new ways of working (Wong & Crowe, 2024). The integration of AI offers the potential to expand the functional capabilities of surgical robots, leveraging multimodal data from sensors, kinematic input and intraoperative imaging to support clinical decision-making and procedural accuracy.

Despite significant technological progress in robotic surgery, challenges persist, particularly in managing unstructured data, nuanced intraoperative decisions and unpredictable clinical scenarios. AI models are being developed to address these limitations by improving image segmentation, integrating multimodal data sources and enhancing image classification accuracy in real time (Liu et al., 2024b). In addition, while a limited number of approved systems dominates the current surgical robotics landscape, advances in integrating AI and surgical robotics may increase the number of robotic platforms (Marcus et al., 2024). Levels of autonomy in robotic surgery vary depending on the degree of human oversight, ranging from direct manual control to shared or autonomous task execution (Iftikhar et al., 2024). A recent scoping review of intraoperative applications of AI in robotic surgery found that the field is still nascent, and most applications have a low level of autonomy (Vasey et al., 2023).

Evidence synthesized by Knudsen et al. (2024) indicates that AI is enhancing surgical robotics by improving the visibility and accuracy of imaging as well as by automating certain steps. For instance, CNNs and transformer-based models have been applied to improve the quality of surgical video by removing visual obstructions such as smoke, thereby improving intraoperative visibility (Wang et al., 2023a). In addition, deep learning models are being developed to improve tissue recognition, which can allow for more accurate image segmentation of anatomical structures (Kumazu et al., 2021), as well as real-time tracking and identification of surgical instruments (Ping et al., 2023). Additionally, AI can contribute to task automation by defining intraoperative constraints and optimizing the initiation and termination of specific actions (Saeidi et al., 2022). Another area of interest in the field of robotic surgery automation is the use of deep reinforcement learning to train robots and enhance the level of automation, through a combination of expert demonstrations and trial-and-error learning (Esteva et al., 2019; Ma et al., 2020; Qian & Ren, 2025).

Another critical application of AI in robotic surgery is intraoperative performance assessment. Machine learning models can analyse kinematic and video data to provide real-time evaluations of surgical skill, contributing to both clinical feedback and surgical training (Knudsen et al., 2024). Motion-tracking technologies, including those leveraging non-optical sensors, are being used to monitor surgical gestures and behaviours for competency assessment. A systematic review of machine learning applications using data from non-optical motion tracking in surgery (where sensors track the movements of the surgeon's hands, a useful technique where the movement is outside the line of sight) found that these methods can improve precision, facilitate objective assessment and support training frameworks, with most studies focusing on robotic-assisted procedures (Carciumaru et al., 2025).

Should AI be used to support robotic surgery?

Robotics surgery is a rapidly evolving field and existing challenges in evaluating and monitoring these technologies may be further exacerbated with the integration of AI (Marcus et al., 2024). While this integration presents a significant technical promise to transform surgical robotics and health care systems, there are important feasibility and ethical constraints that affect clinical translation, including data availability, skillset, cost and equity.

System readiness and integration considerations

Many AI models, particularly those involving deep learning or reinforcement learning, require large volumes of high-quality, annotated surgical data. However, access to such data is limited due to concerns about patient privacy, variability in procedural standards, fragmented health care systems and the intensive resources required to label surgical videos and instrument movements. Real-time applications also demand robust computational capacity and seamless interoperability with surgical systems, which can be challenging to achieve outside high-resource institutions. Furthermore, the complexity of robotic systems themselves means that a high level of technical skill and continuous training is required, both to operate the systems safely and to interpret AI-assisted feedback.

Governance, ethics and rights considerations

The use of AI in surgical robotics also raises complex ethical and regulatory issues, particularly around safety, accountability and equitable use. While AI can enhance decision-making and surgical accuracy, it also introduces risks if systems malfunction or if model outputs are misinterpreted. These concerns are magnified in the context of semi-autonomous or autonomous functions, where responsibility for intraoperative errors may be difficult to assign. The black box nature of many AI models further challenges the principles of transparency and explainability, which are critical for clinician trust and regulatory oversight.

Ethical use also requires ensuring that AI models are trained on diverse and representative datasets to avoid reinforcing biases that could compromise patient safety. Informed consent must reflect the involvement of AI in surgical decision-making, including potential limitations and failure modes of the technology. Additionally, there is a need for robust governance structures, as highlighted by recent literature, to support ongoing evaluation and accountability.

2.4 Public health and health policy

AI has the potential to support public health and health policy, offering new ways to understand, manage and forecast population health (Table 2.3). This section explores how AI can enhance the processing of data on health determinants, uncover complex relationships between exposures and outcomes, support population segmentation and risk stratification and forecast disease burden. It also examines the use of AI in generating behavioural insights, strengthening infectious disease surveillance and improving public engagement and communication. Together, these applications illustrate how AI can contribute to more data-driven, responsive and inclusive public health strategies, while also raising important questions about feasibility, equity and trust.

Table 2.3 *AI applications in public health and health policy*

Applications	Key findings	Considerations
Understanding, managing and forecasting population health: enhancing the processing of data on health determinants, improving our understanding of population health, population segmentation and risk stratification, and predicting and forecasting disease burden	AI offers new opportunities to understand and manage population health by integrating diverse data sources and uncovering complex patterns of disease. AI models, particularly LLMs, can increase the ability to process and integrate a wider range of health determinants data sources, such as unstructured clinical notes, or social care records. AI presents new opportunities to explore the complexities of disease patterns, risk factors and causal relationships in epidemiological research. AI techniques such as clustering and risk prediction can support targeted interventions by informing population segmentation and predicting adverse health events at the population level. While still at the early stages, AI models are being explored to forecast the burden of disease, complementing traditional statistical approaches in public health planning	Recent advances in AI can complement established public health methods, but their added value over existing statistical approaches remains unclear and requires further evaluation. The fragmentation of health determinants data across multiple, siloed systems poses a major challenge for AI and, indeed, any other analytic methods. When used to process information related to the broader determinants of health, AI models may reproduce existing biases, potentially reinforcing the very inequities these models could aim to address. Public trust and consent are fundamental considerations when bringing together ever more sources of data collected by different agencies or individuals, and require robust and ethical governance and regulatory frameworks

> > *continues*

Applications	Key findings	Considerations
Behavioural insights	AI tools offer opportunities to change how behavioural data are collected, analysed and used for behavioural support and change. Chatbots and virtual health assistants can be leveraged for behavioural change, such as providing smoking cessation support or promoting physical activity. AI models can process health data from wearables and health monitoring devices to identify patterns, trends and anomalies indicating health concerns. In the future, AI could support analysis of these data at the population level, allowing public health authorities and researchers to track trends, evaluate the impact of policies and interventions, and further our understanding of the complexities of human behaviour	AI tools used in behavioural health require ongoing validation against current national guidance to prevent ineffective or harmful advice. Robust data governance frameworks and investment in cybersecurity are needed to minimize the risk of reidentification and data breaches, which can increase with the linkage of personal data from different sources. Individual-level AI-driven prevention strategies should be considered in the context of broader public health measures and acknowledge potential challenges in shifting responsibility from health systems to individuals
Infectious disease epidemiology	AI is emerging as a powerful tool in infectious disease epidemiology, with its research and evaluation accelerating during the COVID-19 pandemic, and with recent advances showing potential to transform how infectious diseases are monitored and managed. AI applications in this field are at different levels of maturity and have been described across a range of areas, including epidemic forecasting, scenario modelling, understanding mechanisms of disease spread, surveillance, classification of pathogens and identification of the source of outbreaks. Public health institutions are assessing the use of AI to leverage non-traditional data sources, such as social media and online reviews, to complement existing surveillance systems and identify emerging public health threats. Algorithms such as graph neural networks are being explored to model disease transmission, holding promise to offer more timely interventions	While the repetitive nature and data intensity of infectious disease monitoring and forecasting make AI well-suited to these tasks, adoption can be limited by the high level of interconnectedness between multiple systems and institutions. The impact of using AI tools in infectious disease response must be carefully evaluated, particularly in the context of outbreaks or epidemics, as failures can undermine public confidence and trust, reduce adherence to health measures, and negatively impact outcomes. Despite its promise, AI adoption in this field requires careful evaluation, particularly regarding model transparency, data quality and the integration of outputs into public health decision-making

> > *continues*

Applications	Key findings	Considerations
Communications and public engagement	AI offers the opportunity to enhance critical public health communication efforts by enabling more dynamic, data-informed and responsive engagement with diverse populations. Machine learning techniques can inform the segmentation of populations with different levels of risk or need, allowing for more precise and impactful health messaging. AI tools can help analyse public sentiment on social media, helping health authorities to identify concerns and adapt communication strategies in real time. Generative AI technologies can help synthesize public health advice from guidelines, translate messages quickly into multiple languages, and tailor content to different audiences. While AI-generated content is a contributor to the spread of health misinformation and disinformation, AI tools have also been proposed as a means to address these issues by supporting efforts to detect and track health misinformation online	Some of the key risks in using AI, particularly generative AI, for public health communication stem from the technology's inherent limitations, including its risk of producing inaccurate content. LLMs can generate misleading yet convincing information, which may undermine public trust and health outcomes, and their use requires robust safeguards and mitigation strategies. AI-generated messages may fail to reflect individual or cultural contexts, and biases in training data can reduce the inclusivity and effectiveness of public health communication. Careful oversight and appropriate legal and ethical safeguards are required to ensure that AI's role in countering misinformation does not inadvertently infringe on individual rights or limit access to diverse perspectives

AI: artificial intelligence; COVID-19: coronavirus disease; LLM: large language model.

2.4.1 Understanding, managing and forecasting population health

Factors outside the health care system shape population health. These factors are often referred to as the wider or social determinants of health, and are "broadly defined as the conditions in which people are born, grow, live, work and age, and people's access to power, money and resources" (WHO, 2025d). They include factors such as housing, income, education and the physical, social and natural environments that people inhabit (WHO, 2024; WHO, 2025d). Several frameworks illustrate how these determinants influence health and health inequities (Dyar et al., 2022), with the Rainbow Model developed by Dahlgren and Whitehead one of the most widely used (Dahlgren & Whitehead, 2021).

Understanding and quantifying the complex interrelationships among these determinants over time is challenging. There are benefits to adopting more data-driven approaches to understand these determinants better and manage population health (WHO Regional Office for Europe, 2023) and AI offers a potential opportunity to enhance the way we forecast the burden of disease and inform policies and interventions that improve population health and address inequalities. Indeed, while AI's benefits are often seen in efficiency gains, its potential in public health also lies in uncovering previously unknown patterns

or relationships that could lead to better health outcomes. Opportunities include expanding the types of data that can be analysed and providing new analytical approaches, thereby potentially improving our understanding of the drivers of population health, both overall and for groups within them, including the role of intersectionality, where an individual's characteristics reflect several distinct characteristics.

Enhancing the processing of data on health determinants

AI and generative AI have the potential to significantly enhance the processing of vast amounts of data, enabling quicker and more efficient analysis. Many health care databases, such as electronic health records, contain both structured data (e.g. clinical codes) and unstructured data (e.g. clinical notes), a large portion of which remains unused. This information can include insights about family history, risk factors and wider determinants of health, which are fundamental to understanding population health status and can support efforts to identify the needs of defined populations. To address this, NLP techniques and LLMs are increasingly being applied or evaluated to extract and process relevant information about the wider determinants of health from electronic health records (Patra et al., 2021) and multimodal data, including images, scanned documents and clinical notes (Liu et al., 2023a). These AI models can convert unstructured data into structured categories, making it accessible for further analysis.

While some health care systems already capture information on the wider determinants of health, such as postcodes and thus characteristics of the area of residence (Fairley et al., 2011), AI tools can help identify and process additional, richer information that may currently be hidden in clinical notes. Furthermore, these tools could be applied to other datasets beyond the health care system. Databases that link health care and social care data in some places enable the integration of health status information with other determinants of health. For example, the Discover Dataset in north-west London, United Kingdom, links primary care, secondary, acute, mental health, community health and social care records (Discover-NOW, 2020). AI tools, such as multimodal AI models, could help further harmonize, standardize and streamline the integration of data from clinical and social care, bringing insights from unstructured data sources. In the future, these models could also help with the integration of data from diverse sources, including environmental monitoring, such as air quality measurements from sensors and behavioural data from wearables. An existing example, from Germany, is the creation of high-resolution models to project heat-related mortality (Wang et al., 2024).

By creating datasets that integrate health and wider determinants of health data, AI can enable complex predictive analytics to forecast disease burden or

identify areas at higher risk of worse health outcomes, accounting for more of the complexity of population and individual health. Therefore, by improving access to and integration of a wider range of data sources, AI tools have the potential to enhance our understanding of health determinants and inform decision-making in public health.

Improving our understanding of population health

AI can transform our understanding of the complex relationships between exposures or risk factors and health outcomes. This task has traditionally been addressed by a range of experimental and non-experimental methods, including natural and randomized experiments as well as cohort and case-control studies. While these methods can identify associations in the populations included, they may be limited in their generalizability to others (external validity) (Britton et al., 1999) and they can be limited in their ability to explore non-linear and dynamic associations that involve feedback loops. In contrast, AI, particularly through unsupervised learning techniques, can help uncover previously unknown, non-linear relationships by processing complex data without preconceived assumptions.

The opportunities and limitations of using machine learning techniques in the field of epidemiology have been widely reported. As described in section 2.3.1, many applications focus on using these techniques to predict the likelihood or risk of an outcome given a set of predictor variables (Wiemken & Kelley, 2020), such as the likelihood of developing a chronic condition given a set of risk factors. For example, researchers at the Robert Koch Institute, Germany, have demonstrated that a machine learning model can predict the mental health of adolescents between 3 and 5 years later (Stuke et al., 2025).

Other important applications include the use of AI to better understand the relationships between exposures and health outcomes. In this context, AI presents new opportunities to investigate the complexities of disease patterns, risk factors and causal relationships in epidemiological research, although caution is needed as causation is assessed not just by associations but by other criteria and, especially, biological plausibility (Hill, 1965).

Understanding the evolution of health and disease over time is highly complex, especially in ageing populations in which many individuals experience multimorbidity, with multiple conditions interacting with each other. Unsupervised machine learning methods, such as clustering approaches, are valuable for identifying patterns in multimorbidity and can enhance our understanding of how these conditions interact (Delord et al., 2024). Moreover, combining traditional machine learning techniques with generative deep learning methods

can provide deeper insights into this complexity. For instance, researchers in Denmark trained and evaluated a model using a dataset that included the entire adult population living with cardiovascular disease. This approach employed variational autoencoders, a type of generative AI model, alongside *k*-means clustering (see Box 1.1), an unsupervised machine learning technique, to identify and track the evolving patterns of multimorbidity over time. By integrating these techniques, the study was able to uncover latent patterns in the progression of multiple conditions, offering a more nuanced understanding of multimorbidity (Holm et al., 2025).

AI is increasingly being applied to identify risk factors, such as in life-course analysis and mortality studies. Life-course epidemiology examines how various exposures throughout an individual's lifetime influence health outcomes. A study evaluating the performance of several deep learning architectures for longitudinal analysis of life-course data found that these models were able to identify dynamic relationships that traditional statistical and machine learning methods could not. However, recalling the concern about biological plausibility noted above, the authors described findings inconsistent with established causal relationships, causing them to call for "alternative XAI techniques that better align with epidemiological principles" (Coupland et al., 2025). In a separate study, researchers applied machine learning techniques to health survey data to identify risk factors for all-cause mortality, using XAI to interpret the results and pinpoint both risk and protective factors (Agogo & Mwambi, 2025). Although these applications have several limitations, they highlight the potential of AI and XAI to address complex epidemiological questions and enhance our understanding of health risks over time.

Causal inference, determining whether a factor leads to a particular exposure, is fundamentally an epidemiological question requiring a combination of empirical observation and reasoning. It is therefore not reasonable to expect one tool to be able to do everything at the present time. However, a review summarizing the integration of AI into causal research in epidemiology described ways in which these tools could fill gaps in the causal roadmap; for example, by using generative models to synthesize research to develop initial hypotheses and identify knowledge gaps; employing AI techniques to create causal structures from data and estimating statistical parameters; and using generative AI to help disseminate findings (Matthay et al., 2025). Other authors have described the use of AI as a copilot to generate causal evidence, supporting researchers in areas such as translating contextual knowledge into assumptions and facilitating the causal interpretation of results (Petersen et al., 2024). Thus, while the limitations of AI tools in differentiating association and causation are important (Sung & Hopper, 2023), there is potential for deep learning approaches and causal

machine learning to help understand these relationships and advance public health research (Feuerriegel et al., 2024; Lagemann et al., 2023).

Population segmentation and risk stratification

A comprehensive understanding of population health requires information not just on aggregate levels of health, disease and risk factors but also on their distributions and, especially, how these are patterned by individual and group characteristics. This process of segmentation involves classifying the population into groups or cohorts which have similar characteristics (e.g. age, sex, ethnicity). Unsupervised machine learning techniques, particularly *k*-means clustering, are commonly used for segmentation in health contexts (Liu et al., 2023b). These algorithms divide the unlabelled training dataset into *k* different clusters of data points which are near each other (Goodfellow et al., 2016). For example, this approach has been used to understand health care disparities in the USA, identifying complex clusters that include multiple characteristics involving both demographic and health care access information (Bowser et al., 2024). Thus, examples of the clusters generated include "rural, poorest, least educated, with lowest health access and health infrastructure, with the lowest life expectancy and least immigration" and "suburban, average poverty, education, insurance and health infrastructure, relatively low immigration". This extends traditional clustering analyses, such as "A Classification of Residential Neighbourhoods" (ACORN), that are derived using conventional principal component analyses (Morgan & Chinn, 1983).

Risk stratification models can be used as part of the segmentation analysis to predict the risk of adverse events. At the population level, this can help to understand patterns and distribution of risk across a population (NHS England, 2018), which can support prioritization of interventions and inform resource allocation. Similar to the individual risk prediction methods described in section 2.3.1, regression analysis and other statistical techniques have been traditionally used for risk prediction, but machine learning approaches can also be applied. In the USA, an ensemble machine learning model was developed and validated to predict neighbourhood-level risk of drug overdose as a tool to guide allocation of resources and inform community-level prevention measures by using historical data from Rhode Island (Krieger et al., 2025). The team also developed an online mapping dashboard to inform outreach efforts (Allen et al., 2024). However, in contrast to the clinical and patient-level risk prediction applications of machine learning and deep learning models, applications at the population level have received much less attention (Bowe et al., 2023).

Predicting and forecasting disease burden

An important part of managing health at the population level is the ability to predict the disease burden to inform public health policy and resource allocation. Advanced AI approaches offer the potential to transform how different sources of clinical and non-clinical data are monitored and analysed to predict future population health outcomes. Forecasting noncommunicable disease trends and long-term population health outcomes has traditionally relied on advanced statistical methods (GBD 2021 Forecasting Collaborators, 2024), such as Bayesian meta-regression and spatiotemporal modelling, which, while highly robust and interpretable, are not typically classified as AI techniques. However, recent developments have begun to explore the potential of machine learning in this space. For example, deep learning models have been used to forecast the incidence of depression at a population level, demonstrating comparable or superior performance to traditional time-series models under certain conditions, such as during societal disruptions (Yang et al., 2025a). Other approaches focus on using machine learning techniques to predict demand for health care services by examining the influence of variables such as demographic and social factors, access to services and health status (Orhan & Kurutkan, 2025). Although implementation of these applications is, so far, limited, these examples underscore the emerging role of AI in complementing traditional forecasting approaches to support data-driven public health planning.

Should AI be used to understand, manage and forecast population health?

While AI holds considerable promise to strengthen approaches to population health through more accurate and efficient access to data, understanding of the drivers of disease, how disease is distributed within populations and disease forecasting methods, this application is still in the early stages of development and implementation in most public health systems. Many existing advanced statistical techniques used in public health are well-established, methodologically robust and well-suited to contexts where transparency, comparability and reproducibility are critical. As such, recent advances in AI can be viewed as complementary tools that can enhance existing analytical methods and lead to new approaches, but there is not yet clarity on their advantages, if any, over existing solutions (Kraemer et al., 2025). Below, we consider four issues: data fragmentation, resources, biases and trust.

System readiness and integration considerations

The fragmentation of health and social data across multiple, siloed systems poses a major challenge for AI and, indeed, any other analytic methods. Effective AI-based approaches to manage population health rely on the integration of diverse datasets,

including clinical records, public health surveillance, pharmacy use, housing status and environmental exposures. However, many public health agencies and health systems lack the technical infrastructure, governance frameworks and interoperability standards required to link and analyse these data at scale securely. Hence, the use of AI in this field faces logistic challenges related to data standardization, missing data and the external validity of models trained in specific populations or health systems but applied elsewhere.

Moreover, while advanced AI models can enhance predictive accuracy and handle complex data, they often require significant computational resources, technical expertise and maintenance, which can pose feasibility challenges and increase implementation costs. In many public health contexts, particularly where there is limited infrastructure, traditional statistical methods remain more cost-effective, interpretable and aligned with existing institutional capacity. The marginal gains in performance offered by AI models may not always outweigh the additional resource needs, highlighting the need for careful cost–benefit assessments when choosing among methodological approaches that provide a better understanding of population health.

Governance, ethics and rights considerations

Biases in LLMs and generative AI models have been extensively reported. If using these models to extract information related to the broader determinants of health from unstructured data in electronic health records, even when attempting to understand the drivers of population health inequities better, there is a risk that the extracted information is also biased, further exacerbating these inequalities. A recent study explored different LLMs to extract data on social determinants of health, such as employment status, housing issues, transportation issues, parental status and social support from clinical texts in electronic health records (Guevara et al., 2024). These models were trained with augmented datasets, which included synthetic data that increased the mentions of certain phrases and, especially, those that were less common in the original data. However, the authors noted how the content of the synthetic data impacted model performance, varying, for example, when different genders were included. This study further adds to the increasing evidence that biases can occur at multiple stages of the AI implementation pipeline. This also highlights the complexity of rolling this out in an equitable manner, as with other types of technology, underscoring the need for checkpoints and evaluation of biases across the technology life-cycle.

Another fundamental consideration when bringing together ever more sources of data collected by different agencies or individuals relates to population trust and consent. In addition to the legal and governance frameworks and ethical safeguards required to guide the data collection, storage and processing pipeline,

there are important questions about whether the public will consent to the transfer of their data to other organizations, whether within the public sector, such as social services, or to commercial companies, such as data from wearables, to be processed for public health purposes. As an example, data from the United Kingdom reported in The Health Foundation's 2024 annual attitudes survey (Binesmael et al., 2024) found that the willingness to have some of their data for AI development was high, but this varied by type of data. For example, while 58% of the public were happy to share information about the medicines they are taking or have taken with those developing AI systems, this went down to 47% for information collected by phone or wearables. Importantly, this varied by socioeconomic groups. People in the lowest socioeconomic groups were significantly less likely to support the use of their health data for AI development compared to the other groups. This highlights another area in AI development and adoption that raises important equity considerations. If future AI models that can help us understand population health are developed only with data from groups with higher incomes, the less affluent will again be underrepresented, contributing to widening inequalities.

2.4.2 Behavioural insights

Behaviour plays a key role in shaping individuals' health outcomes and is influenced by a variety of factors. Some are deeply embedded in the broader social, economic and environmental context in which people live, as highlighted in the discussion of the wider determinants of health in section 2.4.1. Tobacco use, unhealthy diet, alcohol consumption, and lack of exercise are the four leading behavioural causes of years of life lost in the United Kingdom, as in many other countries, and are leading causes of health inequalities (Marteau et al., 2021). Research on behavioural and social sciences seeks to understand the underlying drivers of behaviour to inform public health policies and services (WHO, 2025a), but designing and evaluating interventions to change behaviour remains a complex challenge. In this context, AI tools offer new opportunities to leverage behavioural science insights and data from wearable and health monitoring devices to provide real-time monitoring and personalized advice.

As discussed in section 2.3.3, AI is transforming how chatbots and virtual health assistants can provide support for mental health and chronic disease management. In the context of health promotion, these tools also present significant opportunities for facilitating behavioural change, such as providing smoking cessation support or promoting physical activity. Several AI-based chatbots have been developed to support behavioural change, particularly in the areas of healthy lifestyles and smoking cessation (Aggarwal et al., 2023). For example, WHO introduced SARAH, a digital health promoter which leverages

generative AI to provide tips in multiple languages on making healthier choices (WHO, 2025c). Additionally, smartphone smoking cessation applications have been developed to tailor messages based on data about factors triggering previous smoking episodes, enabling real-time support to manage cravings (Naughton et al., 2023).

Wearable devices, such as smartwatches and fitness trackers, can collect a wide range of physiological data, including heart rate, oxygen saturation, temperature and movement. A systematic review concluded that activity trackers appear to be effective at increasing physical activity in a variety of age groups and clinical and non-clinical populations, and that the benefit was clinically important and sustained over time (Ferguson et al., 2022). AI models can process health data from these devices to identify patterns, trends and anomalies indicating health concerns. For instance, wearable AI-enabled devices can monitor heart rhythms and detect irregularities, such as atrial fibrillation, which might go unnoticed. While not a diagnostic device, the Apple Watch and similar devices can support in tracking atrial fibrillation history, highlighting AI's potential role in complementing traditional diagnostic tools (Shahid et al., 2025). In addition, AI can enhance wearable's ability to provide personalized health insights. AI models can deliver tailored recommendations for improving physical activity, diet and sleep by analysing user-specific data over time. For example, fitness trackers use AI to recommend exercise routines based on an individual's activity history and fitness goals. Similarly, AI-powered devices can analyse sleep, activity and diet patterns and offer suggestions that can help users adopt healthier behaviours (Oura, 2025).

AI can be used to design and personalize behavioural interventions that promote health by analysing individual patterns, preferences and risk factors. For example, in smoking cessation programmes, machine learning models can identify moments of high relapse risk based on user-reported data and sensor inputs, triggering timely motivational messages or support prompts (Perski et al., 2024). AI has also been used to optimize mental health interventions by adapting cognitive behavioural therapy modules to individual user responses (Thieme et al., 2023). NLP can analyse patient conversations or journal entries to detect emotional states and suggest coping strategies (Sarker et al., 2024), while reinforcement learning models can continuously refine intervention strategies based on user feedback and behavioural outcomes (Jayaraman et al., 2024). These approaches may enable scalable, adaptive and data-driven health promotion strategies that are more responsive to individual needs.

In summary, AI tools offer opportunities to change the way in which behavioural data are collected, analysed and used for behavioural support and change. In the future, these data could be further aggregated and analysed at the population

level, allowing public health authorities and researchers to track trends, evaluate the impact of policies and interventions, and further our understanding of the complexities of human behaviour.

Should AI be used to support and understand health behaviour?

Beyond the equity and bias considerations highlighted in other applications, the use of AI in behavioural health interventions raises particular challenges in these areas, such as ensuring quality of behavioural support tools, data governance, safeguarding the public from the risks of manipulation and considering the trade-offs between individual and population-level prevention approaches.

System readiness and integration considerations

Ensuring the quality of the insights and recommendations generated by AI tools is a critical challenge. The efficacy and accuracy of these tools must be evaluated and validated against up-to-date clinical and public health guidance. Without standardized evaluation methods, there is a risk of ineffective or harmful advice, especially for vulnerable populations. Regular updates and validation are necessary to maintain the effectiveness and relevance of AI-driven interventions.

Governance, ethics and rights considerations

Data governance is a major concern, especially with the collection, storage and sharing of sensitive health data. AI tools in this area often rely on personal data collected from wearables and apps, raising concerns about privacy, informed consent and ownership. Combining personal health data with other datasets increases the risk of reidentification and stigmatization, while growing reliance on these systems and the breadth of data they hold makes them attractive targets for data theft and ransom. Users need clarity on how their data are used, and strong safeguards are required to protect against breaches. With many AI tools developed by private companies, concerns about data being commercialized without consent, as well as barriers to leveraging this data for research purposes, are significant.

AI tools in behavioural applications also raise concerns about manipulation, particularly when personalized nudging techniques are used. AI systems may exploit users' vulnerabilities, encouraging behaviours that benefit commercial interests rather than promote positive health outcomes. To prevent this, the development of AI tools must adhere to appropriate legal and ethical frameworks, with transparency and user control in mind. Clear regulatory standards are needed to safeguard users from manipulation and ensure that AI supports informed, autonomous decisions.

Finally, and related to the discussion in section 2.3.2 on precision medicine application, it is important to consider the potential tension in precision

prevention approaches between empowering individuals with personalized insights and shifting responsibility for health from public systems to individuals. Individual-level interventions should be viewed in the context of broader population-level prevention strategies and the wider determinants of health that influence behaviour beyond personal choice.

In conclusion, while AI has great potential in behavioural health, it will be essential to address challenges in data governance, quality assurance and the risk of manipulation

2.4.3 Infectious disease epidemiology

One of the most promising uses of AI in public health research and practice is in the field of infectious disease epidemiology. The study and evaluation of AI tools for epidemics accelerated during the COVID-19 pandemic (Chen & See, 2020; Wang et al., 2021) and recent advances in AI show how infectious disease monitoring and response can be further transformed (Gomez et al., 2025). AI approaches have been described across a range of areas, including epidemic forecasting, scenario modelling, understanding mechanisms of disease spread, surveillance, classification of pathogens and identification of the source of outbreaks (Brownstein et al., 2023; Kraemer et al., 2025). Each of these applications is at a different level of maturity, with epidemic detection and forecasting being one of the most advanced (Kraemer et al., 2025).

The potential of AI in infectious disease surveillance is particularly great in relation to its ability to collect and analyse unstructured data from multiple sources, such as health records, social media and websites. Although the impact has been limited so far, public health institutions are implementing promising applications. For example, the United States Centers for Disease Control has been able to accelerate response and prevention of Legionnaire's disease outbreaks by identifying cooling towers from aerial imagery using deep learning (CDC, 2025). Separately, the United Kingdom Health Security Agency has evaluated the use of LLMs to detect foodborne gastrointestinal illness from analysis of online reviews of restaurants, screening for key terms related to gastrointestinal symptoms and specific types of food (Laurence et al., 2025). In Australia, researchers explored the use of different algorithms to analyse Twitter (now known as X) posts for the early detection of acute disease events, such as thunderstorm asthma (Joshi et al., 2020). German researchers have augmented wastewater-based epidemiology with machine learning, which enables real-time monitoring of pathogen prevalence at the population level, offering a scalable and non-invasive surveillance method (Aßmann et al., 2025). These examples show that AI can leverage data from diverse data sources to complement existing surveillance systems.

AI has significant potential for analysing vast datasets that can enhance our understanding of infectious disease transmission mechanisms. Traditional epidemiological models have limitations in terms of the complexity and scale of data they can process, and machine learning and deep learning approaches can address some of these limitations. A recent review of the integration of AI with infectious disease mechanistic modelling identified several applications spanning 26 different infections, with the majority focusing on COVID-19 (Ye et al., 2025), and described the opportunities and challenges of this modelling. The former includes enhanced forecasting accuracy from analysing complex data patterns, improved model calibration and parameter inference, aligning simulations more closely with real-world disease dynamics, optimization of intervention strategies by simulating a range of scenarios and evaluating their outcomes, and integrating diverse datasets, from social media to clinical records, to create more comprehensive models. The latter includes the need for interdisciplinary collaboration between epidemiologists and AI experts, difficulty in interpreting results (especially with deep learning models that often function as black boxes), the need for high-quality data and the requirement to include domain expertise to ensure that models reflect biological and social realities.

One notable example of such interdisciplinary cooperation concerns the prediction of how climate change will alter the epidemiology of infectious diseases, particularly vector-borne illnesses. For example, machine learning has been used to forecast future habitat suitability for *Aedes aegypti* and *Aedes albopictus*, the primary vectors of dengue and other arboviruses, under various climate scenarios (Siddiqui et al., 2024). This research suggests significant geographical shifts in vector distribution, with implications for disease risk in previously unaffected regions. Other research has used machine learning to model dengue fever incidence by integrating climatic and socioeconomic variables, demonstrating how AI can disentangle complex interactions between environmental change and human vulnerability (Siabi et al., 2024).

Several studies have highlighted the potential of graph neural networks to improve the understanding and forecasting of infectious disease transmission (Kraemer et al., 2025). Graph neural networks are a type of deep learning architecture designed to handle graph-structured data, which consists of nodes (e.g. individuals) and edges (e.g. relationships between individuals, such as contacts). These types of structure emerge naturally in infectious diseases, for example, in the form of disease spread through networks of individuals. Recent studies have demonstrated the applications of graph neural networks for epidemic and pandemic forecasting (Cao et al., 2023; Panagopoulos et al., 2021). By modelling the dynamic and interconnected nature of disease transmission, AI holds great promise for enhancing epidemic predictions and informing more effective public health strategies.

Should AI be used in infectious disease epidemiology?

Surveillance systems typically rely on the extensive, routine collection and analysis of data from a wide range of sources, such as laboratory results and health records. These systems involve highly repetitive tasks and the processing of large volumes of data, activities that AI is particularly suitable for. However, there are several challenges to consider such as data quality and availability, fragmentation between data systems, public trust and ethical concerns.

System readiness and integration considerations

Surveillance data alone can suffer from gaps, biases and time delays, which can hinder timely decision-making in public health. AI tools applied to infectious disease forecasting and modelling offer great potential to address these challenges due to their ability to process complex, non-linear data and adjust predictions in real time, particularly given the repetitive nature of forecasting that requires continuous updates to models. While AI can effectively model these complex systems, certain tasks, such as predicting future human behaviour or the emergence of novel variants, remain challenging, although there are some promising applications. Thus, researchers at Harvard Medical School, USA, developed a tool, EVEscape, offering the potential for identifying viral mutations with increased risk of giving rise to variants of concern (Thadani et al., 2023). Although trained on the SARS-CoV-2 virus, the authors argued that it could be used with other viruses. A Korean group used a portfolio of learning tools to predict the emergence of new SARS-CoV-2 lineages (Choi et al., 2024). Yet, notwithstanding these examples, the challenges involved may limit the extent to which full automation is possible.

The use of surveillance and forecasting data to inform public health decisions often relies on a high degree of interconnectedness between different systems (such as public health authorities, health care organizations, laboratories and academic institutions), which can create challenges for the adoption of AI. While AI can effectively analyse complex relationships within data, the risk of fragmentation arises when AI systems handle interconnected tasks across multiple domains or stakeholders. For instance, if disease modelling outputs are not properly integrated with health care systems, resource allocation or emergency response plans, the overall system's efficiency may be compromised due to coordination failures. The COVID-19 pandemic provided many examples, such as the failure by those modelling spread in care homes to appreciate the role played by staff who worked across multiple sites.

Some of the most critical considerations relate to trust (McKee et al., 2024). This encompasses the trust that policy-makers place in the models when making decisions, as well as the public's trust in the reliability of these systems, which directly influences adherence to public health measures (Grah et al., 2025).

The failure to detect outbreaks in a timely manner or to accurately predict the impact of an infectious disease can have severe consequences. For example, introducing COVID-19 restrictions in the United Kingdom even a week earlier could have saved many thousands of lives (Arnold et al., 2022). This highlights the importance of thoroughly piloting and evaluating models before they are widely adopted. The potential for failure in these systems is significant, with risks including negative impacts on health outcomes, misallocated resources and erosion of public trust. Similarly, during the acute response to infectious diseases or other health threats, public trust is essential if recommendations are to be followed and crucial information about contacts, sources of infection and potential routes of transmission is shared with public health authorities (Reicher, 2024). Therefore, the impact of using AI tools in these contexts must be carefully evaluated to ensure that trust is maintained and information is not lost, as any breakdown in trust could pose a high risk to public health.

Governance, ethics and rights considerations

Ethical concerns centre on fairness and data privacy. AI models can perpetuate inequities if trained on unrepresentative data, leading to biased predictions and unequal health outcomes, particularly for vulnerable populations. Additionally, AI systems must comply with privacy regulations to safeguard sensitive health information. Transparency is also essential for ensuring AI-driven decisions are understood and trusted by health professionals, as unclear models can hinder their adoption and effectiveness. Accountability is critical, as errors in AI predictions can have serious consequences, and systems must be in place to address these failures responsibly.

2.4.4 Communications and public engagement

Effective health communication with the public is a core public health responsibility (Rimal & Lapinski, 2009). It involves three main steps: understanding the need for information, identifying mis- and disinformation and countering it, and tailoring messages to different target groups (recognizing that these may change, especially between pandemic and non-pandemic situations) (Wang et al., 2022).

AI can contribute to the first step, understanding public sentiments about public health issues, policies or interventions. Social media sentiment analysis can provide valuable insights to inform the content and timing of communication campaigns, typically by categorizing emotions within text into positive, negative or neutral sentiments of social media posts (Cambria et al., 2017). AI models can be used to both classify these sentiments and to group them into relevant topics. During the COVID-19 pandemic, several studies applied machine learning and deep learning techniques to track shifts in public sentiment towards key events, such as reactions to lockdowns and social distancing measures (Jalil et al., 2021;

Valarmathi et al., 2024). Frameworks have also been proposed that combine topic modelling and sentiment analysis over time to assess public responses to health emergencies (Aldosery et al., 2024). By monitoring online discussions, public health practitioners can complement other types of public surveys to gauge public opinion, identify emerging concerns and tailor messages accordingly, ultimately enhancing the effectiveness and impact of communication strategies.

The second step involves responding to the unprecedented spread of misinformation and disinformation. This is now driven partly by AI, which risks worsening health outcomes, widening inequalities and eroding trust in the system. For example, false claims about vaccine side-effects are playing a role in the rise of global vaccine hesitancy, contributing to dangerous outcomes, such as the resurgence of preventable diseases like measles (Hotez et al., 2020). The COVID-19 pandemic marked a turning point in this process. However, AI tools have also been proposed as a means to address these issues on online platforms (Lancet, 2025). While many initiatives already exist to counter false health messaging through fact-checking and providing evidence-based information, significant challenges remain in managing the volume and speed at which such information is generated and disseminated. Moreover, countering false information extends beyond fact-checking, as algorithms themselves influence where people's attention is directed. AI-powered tools can automate the detection and classification of false information and facilitate analysis to assess and compare the spread of misinformation across different channels (Pilati & Venturini, 2025). Using real-world data from the COVID-19 pandemic, AI approaches have been proposed for detecting and predicting the spread of misinformation (Lu et al., 2024a). One role that is likely to become increasingly important is the use of AI to understand the siloed nature of social media whereby those using different platforms are exposed to very different messages. One way this has been done has been to establish what is termed a social listening facility. Social listening is an activity defined as "monitoring the understanding, questions, concerns, information voids, narratives, misinformation, and disinformation that circulate in both web-based and offline environments" (Boender et al., 2023). This is being facilitated, in the EU, by the obligation of platforms to provide data to researchers (Wehrli et al., 2024). In summary, while AI presents significant opportunities to enhance efforts in countering misinformation and disinformation, there is a pressing need for more coordinated approaches to leverage its full potential for public health purposes.

The final step, tailoring messages to suit the diverse needs, contexts and differences of understanding of disease across communities, is especially complex but essential to ensure that the information is relevant and effective (Panteli et al., 2025). The potential for better-targeted messages has increased enormously and such

techniques are used extensively; for example, when customers of companies such as Amazon or Netflix see recommendations based on information held on them (Beer, 2013). In public health, population segmentation enables the crafting of tailored health communications with different levels of risk or need. As discussed in section 2.4.1, unsupervised machine learning techniques can be leveraged for population segmentation. For instance, in a mobile health communication programme focused on reproductive, maternal, neonatal and child health, researchers employed machine learning methods to segment the beneficiary population into distinct clusters based on sociodemographic data and phone access and usage. The researchers concluded that digital communication interventions would benefit from a more differentiated design and implementation to maximize the impact of the communication efforts (Bashingwa et al., 2023).

LLMs and other generative AI technologies also offer opportunities to synthesize public health advice from guidelines and translate them quickly into multiple languages, with other aspects of content, such as images, also designed in ways that are appropriate to different audiences. At the individual level, platforms such as WHO's health-promoting chatbot, SARAH, use generative AI to provide evidence-based advice on healthy living in several languages (WHO, 2025c). At the mass communication level, the integration of generative AI with messaging tools such as WhatsApp or Telegram can be leveraged by government bodies for communication during public health emergencies. These systems could involve specialized AI agents working together to provide appropriately targeted information on national health guidelines using specialized public health AI models (Schmalzle & Wilcox, 2022). Additionally, LLMs can generate alternative public health messages on a particular topic at speed, enabling rapid testing with different audiences. Some studies comparing AI-generated and human-generated public health messages, such as the use of folic acid during pregnancy, found that AI-generated messages, created through prompt engineering, ranked higher in message quality and clarity (Lim & Schmälzle, 2023). In summary, generative AI models offer multiple possibilities to enhance the speed and quality of communication about health with the public.

In conclusion, AI offers valuable opportunities to enhance public health communication, from segmenting populations and analysing sentiment to tailoring messages and detecting misinformation. By leveraging AI tools, public health efforts can be more targeted and efficient, enabling real-time adjustments to communication strategies and ensuring messages are relevant and appropriate to different communities. However, to fully harness AI's potential, coordinated approaches are needed to address the challenges of misinformation and maximize the impact of communication efforts.

Should AI be used for communications and public engagement in public health?

While AI offers significant opportunities for public health communication and engagement, several important considerations must be addressed, particularly regarding the limitations of generative AI. Beyond concerns about bias in the content generated and the advice provided, there are risks associated with the dissemination of erroneous information, the lack of contextual understanding by LLMs and the potential misuse of AI as a tool for censorship under the guise of combating misinformation.

System readiness and integration considerations

One of the significant risks of using generative AI in public health communications is the potential for the dissemination of inaccurate or misleading information. As described in section 1.3.2, certain types of AI model are known to hallucinate, generating plausible-sounding but inaccurate or misleading responses. This can spread misinformation, erode public trust and lead to harmful health outcomes if people follow incorrect advice. Although mitigation strategies, such as adding disclaimers stating that AI-generated information may be inaccurate and that medical advice should be sought from health care professionals, can help reduce risks, these safeguards may not be enough. As LLMs become more fluent, empathetic and realistic in their responses, they pose an increased risk of being perceived as more trustworthy sources of information (see section 1.3.1).

Governance, ethics and rights considerations

While LLMs can be useful for tailoring public health messages based on known factors, such as demographics or language, they are limited by their inability to access less easily measured variables related to individual context, including attitude to risk, numeracy or literacy. These models are trained on large datasets that reflect general patterns, but they lack the ability to understand the specific circumstances, values or cultural nuances of the individual asking for advice. As a result, the messages they generate may not fully align with the personal needs or cultural backgrounds of different groups. Furthermore, the data that LLMs are trained on often contains gaps, particularly when it comes to less-represented populations. This can lead to messages that are ill-suited, incomplete or even culturally insensitive, undermining the effectiveness of public health communications. Given that AI systems can reinforce the biases present in their training data, these limitations need to be carefully considered when deploying LLMs in diverse public health contexts. In this context, one interesting observation is that people of East Asian heritage are more positive about engaging with chatbots than those of European heritage (Folk et al., 2025), something that has also been invoked when discussing the acceptability of robots in health and social care in Japan.

While AI has the potential to play a vital role in countering health misinformation and disinformation, its use also raises significant concerns about surveillance and the potential for censorship of free speech. The reliance on AI tools to detect and flag false information could give certain institutions, governments or tech companies disproportionate control over what is considered false or misleading. This power could be misused, leading to the suppression of legitimate debate, the stifling of minority voices or the unwarranted censorship of information that does not align with prevailing narratives. While the risks of such abuses may be less pronounced in the health sector compared to other areas, they remain a concern. Inaccurate or biased AI models could inadvertently target discussions or perspectives that challenge the dominant health policies or public health messages, thus limiting open dialogue and undermining trust in both AI and health authorities. Thus, careful oversight and appropriate legal and ethical safeguards are required to ensure that AI's role in countering misinformation does not inadvertently infringe on individual rights or limit access to diverse perspectives.

In summary, while AI holds promise for enhancing public health communication, several critical risks must be considered carefully. These include the potential for disseminating erroneous information due to AI hallucinations, the inability of LLMs to fully understand individual context and cultural nuances, and concerns about surveillance and the potential for censorship. These risks, if not properly managed, could erode trust in public health authorities and have adverse consequences for public health outcomes, underscoring the need for careful oversight and a balanced approach to ensure the responsible use of AI in health communication.

2.5 Research and innovation

AI is transforming the landscape of health research and innovation, particularly in the development of new therapies (Table 2.4, page 84). This section explores how AI is being used to accelerate drug discovery, both by improving predictive capabilities in early-stage research and by leveraging generative models to design novel compounds. It also examines emerging applications in clinical trial processes, where AI is supporting trial design, recruitment and simulation through tools such as digital twins. These innovations offer the potential to reduce costs, improve efficiency and enhance the inclusivity of biomedical research, while also raising important questions about safety, transparency and regulatory oversight.

Table 2.4 *AI applications in research and innovation*

Applications	Key findings	Considerations
Accelerating drug discovery: improving predictive capabilities in early-stage drug discovery and leveraging generative AI to design new drugs	AI is increasingly regarded as a transformative tool to accelerate early-stage drug discovery and candidate design, addressing persistent challenges in the drug development process. Machine learning predictive capabilities can be leveraged in several stages of early drug development, ranging from identifying viable targets, to predicting drug interactions and assessing potential adverse effects. Advances in generative AI are enabling the de novo design of novel drug candidates and protein structures, shifting the focus from prediction to design of new molecules that fulfil a desired function or property	Despite ongoing progress and a growing research pipeline of AI-discovered drugs, none have yet reached the market. The shift from predictive to generative AI models, while promising for de novo molecule and protein design, raises specific concerns around the validity and safety of their outputs. Effective use of AI in drug development is constrained by high data and resource demands, as well as disconnect between early discovery efforts and downstream processes. Many available datasets focus on common diseases and well-characterized biological targets, which may perpetuate existing disparities in biomedical research and limit the generalizability of AI-driven findings
Improving clinical trial processes	AI holds the potential to address longstanding inefficiencies in clinical trials by making them more predictive, adaptive and inclusive. Machine learning models have been developed to improve the efficiency and effectiveness of clinical trials by predicting a range of relevant factors, such as factors influencing trial participation. Digital twins are increasingly being explored for their potential to serve as synthetic control arms, where predicted outcomes in the digital twin counterpart can be compared with those of participants receiving an intervention	Integrating AI into clinical trial workflows requires substantial technical infrastructure, workforce adaptation and clear regulatory pathways to ensure safe and effective deployment. The use of proprietary AI tools in clinical trials raises concerns about transparency, fairness and data privacy, particularly when models lack public scrutiny and rely on sensitive health data

AI: artificial intelligence.

Health research and innovation encompass a broad range of disciplines and activities aimed at improving health outcomes and population well-being. The potential applications for AI are manifold, especially in areas that require analysis of large and complex datasets, such as those with genomic, proteomic and metabolomic information. Examples include identification of disease-associated genetic variants by scanning whole-genome sequencing data (Duong & Solomon, 2025); the use of deep learning models to interpret proteomic data, to identify protein biomarkers for diseases or predict protein structures and interactions; and the use of machine learning to distinguish metabolic signatures that facilitate early diagnosis (Chi et al., 2024). In this section, however, we will focus on one area, the development of new drugs. Specifically, we will explore the use of AI in accelerating early-stage drug discovery and improving the efficiency and effectiveness of clinical trials.

The discovery and development of new drugs is a complex, resource-intensive process that comprises several stages. It typically begins with a pre-discovery phase, where potential biological targets implicated in the disease of interest are identified. This is followed by the drug discovery stage, during which candidate compounds are screened for their ability to interact with these targets. Successful candidates advance to preclinical testing to assess safety and efficacy in vitro and in animal models, followed by clinical trials in humans. The process concludes with regulatory approval and post-marketing surveillance to ensure continued safety and effectiveness (Singh et al., 2023). Despite decades of progress, drug development remains a slow and costly process. In this context, AI is increasingly viewed as a transformative tool that can address some of these challenges and reshape drug development pipelines. It is important to acknowledge, however, that this is not the first time the pharmaceutical sector has invested in AI with high expectations. Many ventures have struggled to deliver meaningful results over the past decade. Nonetheless, recent advances, particularly in generative AI, have led some experts to suggest that the potential of AI in drug discovery warrants renewed attention.

2.5.1 Accelerating drug discovery

A recent review of AI applications in drug development highlights several promising uses of these technologies in early-stage research (Zhang et al., 2025), leveraging both the predictive capabilities of machine learning and the generative capabilities of more recent AI models.

Improving predictive capabilities in early drug discovery

Machine learning approaches offer predictive capabilities that can be leveraged in several stages of early drug development, ranging from identifying viable targets to predicting drug interactions and assessing potential adverse effects.

AI can improve the identification of drug targets by analysing complex biological patterns and relationships that would be difficult to discern through traditional methods. For example, machine learning methods have been proposed to predict cancer genes by integrating genomics and protein–protein interaction networks. This method enabled researchers to identify 165 novel cancer genes that interact with known cancer genes and which offer new opportunities for precision oncology and for identifying biomarkers for other complex conditions (Schulte-Sasse et al., 2021). While, conventionally, drug-candidate screening involves well-established methods, such as high-throughput biochemical experiments and computational methods that predict interactions between the target and the drug candidate, AI offers the potential to enhance their operations using machine learning and deep learning approaches. These include those based on AlphaFold technology, which can predict protein structures and interactions

between different biomolecules using protein sequences or other molecular data as inputs (Abramson et al., 2024; Ru et al., 2024).

Beyond identifying drug–target interactions, AI is increasingly being used in pharmacokinetics to predict the absorption, distribution, metabolism, excretion and toxicity (ADMET) properties of drug candidates. Deep learning approaches are showing particular promise by automatically extracting meaningful features from simple input data rather than requiring manual feature engineering (Zhang et al., 2025). These AI-driven advances are substantially accelerating the identification of promising therapeutic compounds.

Leveraging generative AI to design new drug candidates

In addition to leveraging the predictive capabilities of machine learning, advances in generative AI are enabling the de novo design of novel drug candidates and protein structures, marking a significant shift in therapeutic innovation. Rather than predicting molecular structure or interactions based on a known sequence, generative approaches aim to design new molecules that fulfil a desired function or property, which is subsequently validated experimentally.

Deep learning techniques have already been successfully applied to identify new structures and develop proteins and small molecules (Zhang et al., 2025), with generative chemical language models treating chemical sequences in a similar way to how language models process text. These models learn the chemical sequence language from the textual representation of molecules as inputs, and can be further trained to predict bioactivity. These models support the creation of targeted virtual libraries, streamlining the screening process by focusing on candidates more likely to exhibit desired properties (Moret et al., 2023). Beyond chemical compounds, generative models such as diffusion-based approaches, which work by adding noise to data and then reversing this process to regenerate new data, have been applied to generate novel functional proteins with high experimental success rates, with these methods allowing for the generation of proteins that can interact with specific targets based on minimal molecular specifications (Watson et al., 2023).

In summary, generative models can be integrated into an AI-driven molecular design pipeline, leveraging diverse molecular representations to train models capable of generating novel compounds, which can ultimately accelerate the identification of novel drugs (Zhang et al., 2025). However, this has risks as the chemical entities developed can be used for good or ill.

Should AI be used to accelerate drug discovery?

While the integration of AI into drug discovery offers substantial promise, several ethical and feasibility-related limitations should be considered, including issues around reliability and safety of generative models, data quality and availability, and other feasibility constraints (Zhang et al., 2025).

System readiness and integration considerations

The shift from predictive to generative AI models, while promising for de novo molecule and protein design, raises specific concerns around the validity and safety of their outputs. These models can produce hallucinated structures, molecules that appear novel but are chemically non-viable or synthetically infeasible. Although these regions tend to be labelled as low confidence by the models, rigorous experimental validation and integration with other types of modelling remain necessary to mitigate these risks (Abramson et al., 2024).

From a practical standpoint, the successful application of AI in drug development faces several feasibility challenges. Training sophisticated models requires substantial computational resources and access to large, annotated datasets, both of which can be costly and are often unavailable for novel or neglected therapeutic areas. Furthermore, the research ecosystem remains fragmented, with some approaches focusing heavily on early-stage discovery without incorporating downstream considerations, such as compound synthesis (Zhang et al., 2025). This disconnect can hinder the translation of AI-generated compounds into viable therapeutics and may contribute to the current lack of AI-developed drugs reaching the market.

Governance, ethics and rights considerations

AI models in drug discovery are highly dependent on the quality and representativeness of the data on which they are trained. Many available datasets can be biased towards common diseases and well-characterized biological targets, leading to underrepresentation of rare diseases and complex patient populations, such as those with multimorbidity. This can perpetuate existing disparities in biomedical research and limit the generalizability of AI-driven findings. The lack of knowledge about the potentially multiple mechanisms of action of some drugs can make subsequent steps, such as compound screening and therapeutic optimization, difficult to implement. For example, the full spectrum of mechanisms of action of metformin, a drug taken daily by over 200 million people, is still not fully understood (Foretz et al., 2023). Without deliberate efforts to improve data diversity and accessibility, AI may inadvertently reinforce inequities rather than reduce them (Blanco-Gonzalez et al., 2023).

Despite ongoing progress and a growing research pipeline of AI-discovered drugs, none have yet reached the market, underscoring the gap between early

computational success and real-world therapeutic implementation. As AI continues to evolve within drug discovery, it is imperative that ethical considerations, such as data equity, transparency and accountability, be integrated from the outset.

Dual use

A note of caution is required. The integration of AI into life sciences presents significant risks from what is termed the dual use of knowledge that can be employed for both beneficial and harmful purposes (National Academies of Science Engineering and Medicine, 2025). This is now referred to as AIxBio risk. For example, an AI model designed to accelerate drug discovery by predicting protein structures or simulating molecular interactions can also be repurposed to design toxic compounds or engineered pathogens. For example, for a conference presentation, a commercial company was asked to conduct a demonstration of the potential of AI to create novel toxic molecules similar to the nerve agent VX. Working overnight, it was able to suggest 40 000 substances that included VX and other known chemical agents, but also many other previously unknown ones with toxic potential (Sohn & Stix, 2022). Gene editing technologies, such as CRISPR, can be used to cure genetic diseases or to accelerate the design and synthesis of pathogens by analysing vast datasets on viral and bacterial genomes, protein structures and host–pathogen interactions. This capability could be exploited to engineer organisms with enhanced virulence, resistance to treatment or targeted effects on specific populations.

The risks are increased because of the ability of AI to democratize knowledge and tools. AI-driven platforms can lower the technical barriers to entry, enabling individuals or groups with limited expertise to design harmful biological agents. Open-source models and publicly available genomic databases, while valuable for research, could be misused if combined with AI tools capable of predicting the effects of genetic modifications. AI can also optimize delivery mechanisms for bioweapons, such as aerosol dispersion or vector-based transmission, increasing their effectiveness and stealth. It can also assist in evading detection by simulating how pathogens might behave under different environmental or immunological conditions, complicating early warning systems and response strategies. However, some have argued that these risks should not be exaggerated because, even when candidate molecules are specified, the technical challenges of producing and testing them are considerable (Norwood, 2025).

The dual-use nature of AI in biology, where the same tools used for beneficial research can be repurposed for harm, poses a regulatory challenge. Current oversight mechanisms may not be equipped to monitor or control the rapid pace of AI development and its intersection with synthetic biology. In July 2025, the Nuclear Threat Initiative (2025), a global security organization focused on threats imperilling humanity, issued a statement calling for urgent international

collaboration, technical safeguards and agile governance to prevent the misuse of AI in life sciences, particularly in designing or deploying biological weapons. This statement calls for international cooperation to establish norms for responsible AI use in life sciences, enhancing biosecurity protocols and investing in AI systems that can detect and counteract emerging biological threats.

2.5.2 Improving clinical trial processes

Clinical trials are a key component of biomedical innovation, but continue to face significant challenges. The development of new therapies is notoriously time-consuming, costly and risky, given that only a small number of drug candidates reach the final regulatory approval stages (Kim et al., 2023). Beyond supporting early drug discovery, AI also offers opportunities to reimagine the clinical trial landscape. From predictive modelling and digital twins to LLMs, AI tools are being integrated into several stages of the trial life-cycle, enhancing design, recruitment, monitoring and post-trial analysis.

Machine learning models have been developed to improve the efficiency and effectiveness of clinical trials by predicting a range of relevant factors. One approach leverages data from clinical trial records to create machine learning pipelines capable of predicting early trial termination and identifying associated factors through explainable AI methods (Kavalci & Hartshorn, 2023). Another strategy, the hierarchical interaction network (HINT), focuses on predicting clinical trial outcomes using multimodal data such as drug type, disease target and trial eligibility criteria (Fu et al., 2022). At the patient level, models have been developed to estimate the likelihood of individuals enrolling in trials and to uncover the factors influencing participation, thereby supporting the design of more equitable and inclusive trials (Mehari et al., 2025). Together, these capabilities offer a pathway towards more efficient trial design and improved participant representation.

Next, the integration of digital twins (virtual representation of physical objects, individuals or systems) with AI offers potential for simulation of clinical trials. In health care, digital twins have been defined as a:

> virtual representation of a person which allows dynamic simulation of potential treatment strategy, monitoring and prediction of health trajectory, and early intervention and prevention, based on multi-scale modelling of multimodal data such as clinical, genetic, molecular, environmental, and social factors (Katsoulakis et al., 2024).

In clinical trials, digital twins are increasingly being explored for their potential to serve as synthetic control arms, where predicted outcomes in the digital twin counterpart can be compared with those of participants receiving an intervention,

which can reduce the number of patients needed, increasing the efficiency of the trials (Hutson, 2024). To achieve this, machine learning models can be trained with historical clinical trial and observational study data to help predict individual patient health outcomes in the trial for a specific disease (Bordukova et al., 2024). Beyond simulation, generative AI techniques can also create synthetic data that mimics real patient data while preserving privacy (Kokosi & Harron, 2022). This could be particularly valuable in early-phase research, rare disease trials or when augmenting datasets for underrepresented groups.

Ultimately, LLMs hold significant potential to enhance various aspects of the clinical trial process. LLMs can rapidly synthesize information from prior trials and scientific literature to support more informed and efficient trial design. They can also assist in structuring, standardizing and harmonizing clinical trial data, thereby reducing administrative workload and improving data consistency and quality (Hutson, 2024). Another emerging application of LLMs is recruiting patients to clinical trials. For instance, the TrialGPT framework has been proposed to streamline this process through a three-step approach, following a patient description prompt: retrieving candidate trials from a large database, assessing patient eligibility using relevant criteria, and aggregating the criterion-level predictions to rank the trials. Initial results from this pilot indicate that the model was able to predict eligibility with high accuracy and reduce screening time compared to manual review by clinical experts (Jin et al., 2024).

In summary, AI holds the potential to address longstanding inefficiencies in clinical trials by making them more predictive, adaptive and inclusive. Through the strategic application of machine learning, digital twins and LLMs, trials can be optimized from design through analysis.

Should AI be used to accelerate clinical trials?

While the integration of AI into clinical trials presents exciting opportunities for innovation, it also raises important concerns about ethics and feasibility that must be addressed. As AI technologies begin to reshape trial design, recruitment and data analysis, their adoption must be carefully balanced with safeguards that uphold transparency, equity, privacy and operational viability. Without these considerations, the risk of exacerbating existing challenges, such as disparities in access, regulatory opacity and data misuse, could undermine the improvements that AI seeks to deliver.

System readiness and integration considerations

From a feasibility standpoint, several practical challenges may hinder the integration of AI in clinical trial workflows. High computational demands and infrastructure costs can be prohibitive, especially in low-resource settings or smaller research organizations. Moreover, incorporating AI into existing clinical

trial systems requires not only significant technical adaptation but also workforce training, workflow redesign and clear regulatory pathways. Current frameworks for evaluating AI-powered tools often lag behind technological developments, leading to uncertainty about how these innovations will be assessed by regulatory authorities when used to support trial end-points or replace traditional control arms. Consequently, robust regulatory oversight and clear guidance will be critical to ensuring AI tools are deployed safely and effectively, and that their use is clearly justified in submissions for trial approval or drug registration (NICE, 2025).

Governance, ethics and rights considerations

A key ethical concern relates to transparency and reproducibility, particularly when AI tools, such as predictive models or digital twins, are developed by commercial entities. Many of these tools operate as proprietary systems, limiting public scrutiny of their underlying algorithms, training data and performance across diverse patient populations. This opacity not only constrains the potential for independent validation and generalizability but also raises questions about fairness, especially if models are not representative of minority or underserved groups. Additionally, AI models typically require access to large volumes of sensitive health data, amplifying risks to patient privacy and data security. Ensuring that these data are collected, stored and used in compliance with ethical standards and data protection regulations is essential, particularly given the potential for reidentification from even anonymized datasets.

In conclusion, while AI has the potential to significantly enhance the efficiency and inclusiveness of clinical trials, its ethical and practical limitations must be proactively managed. Ensuring transparency, protecting data privacy, addressing infrastructure gaps and establishing strong regulatory standards are not merely technical concerns, they are essential conditions for building trust and ensuring equitable access to the benefits of AI-driven innovation.

2.6 Operational efficiency

AI is increasingly being used to improve the operational performance of health systems, offering tools to automate routine tasks, optimize resource allocation and accelerate evidence synthesis (Table 2.5, page 92). This section explores how AI can reduce administrative burdens through automation, support more responsive and data-driven decision-making, at both the point of care and the population level, and enhance the speed and scale of evidence generation. These applications are among the most immediately implementable in health care settings, yet they also raise important questions about safety, equity and the integration of AI into existing workflows. Other operational efficiency applications not covered in this book include data integration and management, enterprise resource planning and quality control.

Table 2.5 *AI applications to improve operational efficiency*

Applications	Key findings	Considerations
Automating routine tasks	AI can help reduce administrative burdens in health care by automating routine tasks, potentially improving staff well-being and enhancing patient–clinician communication. Ambient AI scribes, powered by generative AI, are emerging as promising tools for improving clinical documentation and reducing administrative burden, with early evidence on adoption suggesting benefits in terms of efficiency, staff experience and patient engagement. AI tools can also support a range of routine administrative tasks within clinical workflows, including billing and claims processing, as well as the automation of standard patient communications	Generative AI poses risks of producing inaccurate or misleading information, making thorough evaluation, appropriate staff oversight and clear regulatory accountability essential for safe integration into clinical workflows. As ambient AI tools begin to generate summaries and draft clinical recommendations, there is a risk that clinicians may over rely on AI outputs, potentially compromising their critical reasoning. The uneven adoption of AI tools across health care systems and settings raises concerns about the scope to widen health inequalities
Resource allocation and decision-making: improving operations at the point of care and supporting decision-making and prioritization at the population level	AI presents opportunities to streamline a wide range of operational and decision-making processes across the health landscape. Machine learning approaches can support health care organizations in predicting future needs and identifying workflow drivers in complex systems. AI techniques have the potential to support decision-making in priority-setting by helping to navigate their inherent complexity and integrate new forms of data, such as guiding the geographical distribution of resources and supporting development of consensus among stakeholders	The use of AI to guide resource allocation decision-making in health care necessitates a robust data infrastructure and seamless system integration. AI systems using geospatial or environmental data must be governed by strict safeguards to prevent reidentification and protect privacy. AI tools used in policy and participatory decision-making should consider transparency and inclusion, with governance frameworks that ensure diverse voices are represented and digital barriers are addressed
Evidence synthesis	LLMs can significantly accelerate evidence synthesis in health by automating the summarization of large volumes of complex information used in research, clinical guidance and policy development. In systematic reviews, LLMs can enhance workflow efficiency at certain steps, but current evidence does not support its use as a stand-alone or unsupervised alternative to established methods. LLMs offer promising opportunities to improve the speed, accuracy and consistency of health guidance development by automating the review and alignment of complex documentation across agencies and levels	AI tools used in evidence synthesis can produce outputs that appear plausible but are factually incorrect or unsupported by the source data, making expert oversight essential to ensure reliability. The use of AI in evidence synthesis raises concerns about reproducibility, as prompt-based querying lacks the transparency and auditability of traditional methods, underscoring the need for considering explainable AI and adherence to established protocols

LLM: large language model.

2.6.1 Automating routine tasks

The growing burden of documentation and other administrative tasks in the health sector poses a threat to staff well-being, job satisfaction and the quality of human interactions, such as between patients and health professionals. One of the key opportunities that AI offers in the health system is the automation of routine and repetitive tasks, such as note-taking, summarizing or generating standardized letters. By alleviating these administrative pressures, AI tools might not only improve efficiency but also generate value in other ways, such as enhanced patient–clinician communication. The use of generative AI to capture meeting notes and summarize action points is becoming increasingly common in many workplace settings, including health care systems where they are being used for tasks such as real-time transcription during consultations and other routine administrative activities.

Ambient AI scribes, also known as AI scribes, integrate automatic speech recognition with generative AI to transcribe patient–clinician conversations and generate outputs, including transcripts, clinical summaries and structured documentation. While AI technologies for automatic speech recognition, NLP and summarization have been used in health care for some time (van Buchem et al., 2021), recent advances in generative AI have significantly expanded their capabilities. Health systems and providers across different geographies are now piloting and implementing ambient AI scribes tools, supported by emerging guidance on their adoption (NHS England, 2025). Although many settings remain in the early stages of adoption and evaluation, initial evidence points to promising outcomes (Albrecht et al., 2025). A multi-phase study in England found that ambient AI produced higher-quality clinical documentation, reduced consultation time and improved clinician experience compared to the standard model of consultation (Balloch et al., 2024). In the USA, The Permanente Medical Group reported a reduction, compared to non-users of the technology, in the time spent writing notes in each consultation of 0.7 minutes for high users and 0.15 minutes for low users, with mostly positive feedback from both clinicians and patients. Mental health, emergency medicine and primary care were identified as the specialities with the most frequent use (Tierney et al., 2025). Other studies have also pointed to the role of this tool in reducing the mental burden of documentation and increasing the sense of engagement with patients (Duggan et al., 2025). While further evidence is needed to assess the impact of ambient AI across diverse clinical contexts, patient groups and outcomes beyond productivity, its rapid uptake reflects growing recognition of its potential to improve documentation quality and enhance overall care delivery.

In addition to note-taking and summarization, AI tools can also support a range of routine administrative tasks within clinical workflows, including

billing and claims processing, as well as the automation of standard patient communications (National Academy of Medicine, 2025). Ambient AI products can generate outputs such as standard medical letters or other documentation and can also suggest follow-up actions, including scheduling appointments or initiating patient referrals (NHS England, 2025). For example, Epic, an electronic health record software provider, is integrating AI across several aspects of its operations. This includes the use of conversational AI to automatically schedule appointments and handle patient billing enquiries, as well as using AI-driven tools to support the assignment of diagnosis and procedure codes (Epic, 2025). A study assessing the effectiveness and safety of using LLMs to assist with electronic patient messaging reported that the majority of clinicians experienced improved subjective efficiency. The use of LLMs also enhanced the consistency of physician responses and increased the amount of educational information included in patient communications. However, the participating clinicians assessed 7% of the summaries as posing a risk of severe harm and 0.6% as posing a risk of death (Chen et al., 2024b). Most of the harmful responses incorrectly determined or conveyed the acuity of the scenario and recommended action. These examples highlight both the potential and the risks of integrating AI into clinical administrative workflows, particularly when it is incorporated into electronic health systems.

In summary, recent advances in AI are already beginning to transform clinical workflows, offering opportunities to increase efficiency and reallocate time currently spent on repetitive administrative tasks towards activities that enhance the overall care experience. Beyond clinical settings, these tools are also being explored, or offer potential for application, in a range of other person-centred environments related to health and the wider determinants of health, such as social care and acute responses to infectious diseases. To fully realize the potential of AI to improve population health and well-being, it will be important to ensure alignment, interoperability and consistency in the adoption and deployment of AI tools across systems.

Should AI be used to automate routine tasks?

Across the diverse applications of AI in health care, the automation of repetitive and administrative tasks stands out as one of the most promising areas for development. The rapid uptake of ambient AI tools in clinical practice across multiple settings reflects this potential. However, as with any use of AI in health, several important considerations must be addressed. These include the potential for errors or hallucinations to affect patient care, the risk of introducing bias into clinical decision-making and unequal access to and effectiveness of AI tools across populations and settings.

System readiness and integration considerations

One of the most significant risks of using generative AI in health remains the risk of errors or hallucinations, instances where AI generates inaccurate or misleading information. This has direct implications for patient safety and quality of care when communication between patient and clinician or data entry into electronic medical records is concerned. While evidence so far suggests that the risk of harm may be low (Chen et al., 2024b), any risk of a mistake could be unacceptably high in certain clinical contexts. This underscores the importance of a thorough evaluation of how AI tools are integrated into workflows, including determining the optimal level of staff supervision and involvement. Robust and up-to-date regulatory systems are also essential to clarify accountability when errors occur, particularly in complex clinical settings where responsibilities may be difficult to disentangle and where miscommunication or omissions could have serious consequences.

Another emerging concern relates to the influence of AI on clinical judgement. As ambient AI tools begin to generate summaries and draft clinical recommendations, there is a risk that clinicians may over rely on AI outputs, potentially compromising their critical reasoning. This is similar to challenges observed with XAI, where users may defer to algorithmic outputs even when they can make independent assessments, hinting at different types of biases in human–AI collaboration (Rosenbacke et al., 2024b). In a study of responses to patients' messages, those written by LLMs and those where the clinician was assisted by LLMs more closely resembled each other than ones written solely by the clinician (Chen et al., 2024b). The authors interpreted this as the clinicians, assisted by LLMs, deferring to it. Thus, careful monitoring and evaluation of AI tools in clinical practice is essential to uphold clinical standards and ensure patient safety, even when these tools are used to automate repetitive or administrative tasks.

Governance, ethics and rights considerations

The uneven adoption of AI tools across health care systems and settings also raises concerns about the scope to widen health inequalities. Organizations with more advanced digital infrastructure are better positioned to adopt and integrate AI, potentially leading to disparities in the quality and efficiency of care delivery. Within organizations, differences in adoption among clinicians have also been documented (Tierney et al., 2025). Equitable adoption will require an understanding of the relationship between AI usage in different settings and factors such as trust, alignment with clinical workflow and demographics. At the patient level, further disparities may arise. For example, ambient AI systems may perform less accurately with patients who speak with certain accents or dialects and some individuals, particularly those from communities with a lower level of trust in the public or health care system, may feel less comfortable disclosing sensitive information in the presence of AI recording tools. These issues require

ongoing attention to ensure that the deployment of AI does not unintentionally deepen existing inequities in care.

Together, these risks underscore the importance of careful consideration when implementing AI in health care, with a particular focus on context. While the automation of administrative tasks holds considerable promise, realizing its full benefits will require more than technological capability alone. Attention to patient safety, equity, professional integrity, legal and ethical frameworks, and system-level readiness will be crucial in ensuring that AI contributes meaningfully to improving population health and reducing, rather than reinforcing, disparities in care. Importantly, organizations will need to define the goal of integrating AI in clinical processes, assess its impact and reallocate any resources gained in efficiency.

2.6.2 Resource allocation and decision-making

Beyond automating routine tasks such as documentation or summarization, AI also presents opportunities to streamline a wide range of operational and decision-making processes across the health landscape. At the health care organizational level, AI models can enhance resource management and improve patient flow by enabling predictive planning and dynamic scheduling. At the population level, AI can support the optimization and prioritization of service delivery and facilitate coordinated decision-making.

Improving operations at the point of care

Ensuring operational efficiency in health care delivery organizations, such as hospitals and clinics, is crucial for delivering high-quality care and reducing costs, particularly in the context of increasing population demand and health care staff shortages. Despite progress in the digitalization of health care systems in many settings, the adoption of robust tools to support operational management has often been slow, partly due to the complexity of health care operations. Machine learning approaches have been proposed to support health care organizations in two types of operational problem: predicting future needs and identifying workflow drivers in complex systems (Pianykh et al., 2020). We now look at the use of these methods in specific domains of resource management.

Rostering of staff and patient scheduling

AI-driven dynamic scheduling approaches can help match patient demand to health care provider supply, reducing waiting times and improving service delivery. Techniques such as supervised and unsupervised machine learning are used to cluster patients and predict service times, enabling adaptive scheduling (Feng et al., 2024). Additionally, AI tools support patient appointment scheduling to reduce delays and improve flow through hospital departments (Canadian Agency for Drugs and Technologies in Health, 2024).

Forecasting bed capacity and patient flow

Predictive machine learning models can leverage historical and real-time data to forecast hospital admissions and bed occupancy, enabling proactive resource planning and optimizing patient flow (Maleki Varnosfaderani & Forouzanfar, 2024). For instance, real-time data from emergency departments can be used to generate short-term forecasts of admissions by aggregating individual patient-level probabilities of resource use (King et al., 2022). These forecasts help hospitals manage bed capacity within narrow time windows and improve operational responsiveness.

Managing supplies and stockpiles

AI can support smart management of medical equipment and supplies by predicting maintenance needs and anticipating failures, which allows timely interventions and reduces downtime (Zamzam et al., 2023). This predictive capability extends to stockpile management, helping ensure critical resources are available when needed.

Logistics and operational efficiency

Machine learning techniques offer insights into the underlying drivers of hospital inefficiencies. For example, AI has been used to analyse delays in radiology report generation to identify bottlenecks and inform targeted improvements (Pianykh et al., 2020). Ambient AI systems can monitor environmental variables, such as temperature, movement and object location, to optimize space utilization and safety (Haque et al., 2020). Deep learning models also contribute to understanding and managing hospital congestion (Deng et al., 2023), supporting more strategic approaches to workflow design and resource allocation.

Overall, AI offers significant opportunities to optimize resource management in health care settings, particularly by using machine learning to more accurately predict fluctuations in both demand and supply, as well as to uncover underlying patterns.

Supporting decision-making and prioritization at the population level

In addition to applications at the hospital and organizational levels, AI can also play a significant role in supporting resource allocation at the health system and population levels. In these contexts, priority-setting decisions are often guided by a range of ethical, economic and policy frameworks that help determine how resources should be distributed and which groups should be targeted, particularly in resource-constrained environments. Key questions include how decisions are made; the criteria for decision-making, target population groups or geographies; the selection of technology or interventions; and the budgetary impact. AI methods have the potential to support decision-making in priority-setting by helping to navigate AI's inherent complexity and integrating new forms of data

into the decision-making process. Particular contributions to priority-setting at the population level include guiding the geographical distribution of resources and supporting the development of consensus among stakeholders.

Geospatial technologies, such as geographic information systems (GIS), are essential tools for collecting, managing and analysing geographical data to support more informed decision-making in the allocation of health resources, including vaccines. These technologies provide critical insights into both the natural and built environments, such as the spatial distribution of populations, transportation networks and health care infrastructure, which are vital for identifying underserved areas or regions with additional needs (Cunard Chaney & Nagi Mechael, 2020). Geospatial AI (GeoAI) enhances traditional GIS capabilities by integrating spatial data with AI techniques, such as machine learning. This integration allows for more complex modelling, prediction and real-time analysis of spatial health data (Song et al., 2023). In population health, GeoAI offers the opportunity to support more precise, adaptive and equity-oriented decision-making. For example, a framework for vaccine allocation that combines GIS and AI methods was utilized to assess vulnerability factors at the local level and inform vaccine allocation decisions. In addition to leveraging AI to translate data into prioritization criteria, this approach enabled inclusion of equity values in the geospatial vaccination allocation process (Shayegh et al., 2023). Other innovative applications include the optimization of vaccination operations through the use of GIS, unsupervised machine learning and spatial optimization models (Mengüç et al., 2023). Future directions include the development of multimodal GeoAI foundation models capable of processing, generating and integrating diverse data types with geospatial information (Resch et al., 2025). These applications demonstrate how GeoAI can move beyond descriptive mapping to enable real-time, predictive and ethically guided public health resource allocation.

Priority-setting in public health must involve the public and other relevant stakeholders in the decision-making process. These stakeholders, ranging from government agencies and health professionals to civil society and community members, bring diverse perspectives, values and priorities, as well as access to different sources of information. Establishing clear and transparent deliberative processes is therefore critical to ensuring accountability, legitimacy and the effectiveness of public health policy-making. AI offers a range of tools that can support and enhance these processes. As previously noted, NLP and LLMs can help synthesize large volumes of evidence from diverse sources, helping stakeholders quickly develop a shared understanding of complex issues, prevailing themes and proposed interventions. Additionally, AI techniques, such as topic classification and clustering, can be used to extract and group stakeholder inputs by themes or positions, thereby facilitating more structured deliberation and

identifying areas of consensus or disagreement. Participatory AI tools are also being increasingly used to support public engagement and co-creation. For example, open-source AI-driven platforms, such as Pol.is, are being used to support online public deliberation and consensus-building by gathering and analysing participants' conversations (Pol.is, 2025). Together, these applications highlight the potential of AI to strengthen participatory decision-making.

In these ways, AI offers significant opportunities to enhance and innovate resource allocation processes across a wide range of health settings. As discussed in previous sections, machine learning techniques can help predict health care demand at the population level, enabling more proactive and data-driven planning. AI can also contribute to improving priority-setting by generating new insights to guide the distribution of resources, as well as by supporting the decision-making processes themselves.

Should AI be used in resource allocation and decision-making in health?

While AI can enhance efficiency and responsiveness in resource allocation and decision-making across health care settings and populations, it will only be successful if its implementation and adoption is feasible and if it upholds ethical standards related to privacy, transparency and inclusivity.

System readiness and integration considerations

From an operational perspective, the use of AI in health care settings, such as hospitals and clinics, necessitates a robust data infrastructure and seamless system integration. Predictive AI models need access to longitudinal and continuously updated hospital data to capture seasonal trends, long-term workflow patterns and real-time decision points. However, in many contexts, digital infrastructure remains fragmented or underdeveloped, limiting the capacity to generate accurate forecasts or integrate AI tools into routine workflows. Moreover, interoperability between AI tools and existing clinical information systems is often limited, and ensuring that these technologies can support, rather than disrupt, clinical decision-making requires active engagement with health care professionals during the design and deployment phases.

Governance, ethics and rights considerations

In both hospital and population-level applications, AI systems that rely on geospatial data or environmental sensors raise important privacy and data protection concerns. GeoAI tools may use location data, imaging or real-time monitoring of people and environments, data types that, while powerful, can risk reidentification of individuals or facilitate unauthorized surveillance if not properly governed. This is particularly relevant in smaller or marginalized populations, where even anonymized data may be traceable. Hence, ethical use

of such data must be guided by clear safeguards, including data minimization, anonymization and strict access controls that balance utility with individual rights (Haque et al., 2020).

Ethical considerations also extend to the use of AI in policy and participatory decision-making. AI models must be transparent in how they produce their outputs, particularly when used to inform resource allocation or priority-setting. Black box systems that lack explainability can undermine trust and accountability. Furthermore, public participation tools that rely on AI, such as online deliberation platforms, may exclude certain groups due to unequal access to digital tools, low levels of digital literacy or linguistic barriers, thereby potentially reinforcing existing inequities rather than addressing them. To ensure ethical use, AI systems must be designed to promote inclusion and must be accompanied by governance frameworks that enable diverse voices to shape decisions meaningfully.

In summary, while AI has the potential to transform resource allocation and decision-making in public health, its deployment must be grounded in robust technical, ethical and governance foundations. Addressing issues of data quality, privacy, explainability and equity will be critical to ensuring that AI contributes to more inclusive, accountable and effective health systems.

2.6.3 Evidence synthesis

Summarizing and comparing large volumes of information is central to many processes in health. This can include synthesizing scientific evidence for research reports, systematic reviews, policy briefs or guidelines; distilling current guidance to inform clinical decision-making; and ensuring consistency across public health recommendations and protocols. While methodological standards exist to guide many of these processes, they are often labour-intensive, complex and time-consuming. In this context, LLMs, which are well-suited to processing and summarizing extensive textual data, have the potential to increase the speed and scale of evidence synthesis significantly. In health, two specific examples of application are the use of AI in systematic reviews and in consolidating guidance.

Systematic reviews are a cornerstone of evidence-based research, providing a structured approach to synthesizing the scientific literature and informing health policy and clinical decision-making (Mulrow, 1994). Systematic reviews follow several steps, including formulating the research question, designing the search strategy, screening and selecting studies, extracting the information, assessing the quality and bias and synthesizing the evidence. The potential of AI to support various stages of systematic reviews has been extensively documented (Siemens et al., 2025), with studies exploring the use of both generic LLMs (Alshami et al., 2023) and specialized AI products (Cochrane, 2025). A scoping review found

that most LLM applications focused on the stages of literature searching, study selection and data extraction, and that approximately half of the reviewed studies regarded their use as promising (Lieberum et al., 2025). In the search phase, AI can assist in identifying relevant studies using natural language prompts, and support the design of search strategies; however, concerns persist regarding transparency, replicability and completeness (Cochrane, 2025), although this is a rapidly changing field, with continuing improvements in the tools available. A systematic review assessing the use of generative AI for evidence synthesis noted that, while time efficiencies were achieved in the search design phase, a substantial proportion of relevant studies were missed. In the screening, the same review found considerable variability in reported error rates (Clark et al., 2025). The use of LLMs is less established in the text extraction phase (NICE, 2025). Overall, while the use of AI, particularly LLMs, shows considerable promise in accelerating systematic review processes and enhancing workflow efficiency, current evidence does not support its use as a stand-alone or unsupervised alternative to established methods.

Evidence generated through systematic reviews and other scientific methods plays a critical role in informing guidance and recommendations, which are essential for ensuring high-quality clinical and public health advice. In this context, AI offers new opportunities to improve the consistency of health guidance across regional, national and international levels. In many settings, reviewing, updating and consolidating guidance documents remain processes that are mostly manual, time-consuming and prone to error. These activities include tasks such as translating international recommendations into national guidelines and ensuring coherence across documents produced by the same agency or multiple agencies. The United Kingdom Health Security Agency is exploring the use of LLMs to identify potential conflicts between public health guidance recommendations, aiming to support more consistent public messaging, particularly during health emergencies. The approach involves uploading a draft document, automatically searching relevant content from existing agency guidance and flagging any inconsistencies (UK Health Security Agency, 2025). At the international level, WHO's SMART Guidelines initiative, designed to make recommendations standards-based, machine-readable, adaptive, requirement-based and testable, is intended to support the operationalization of guidance (Mehl et al., 2021) and could also facilitate the creation of digital environments in countries where AI tools can achieve greater consistency and efficiency. While these approaches are still in the early stages of implementation, the use of LLMs has the potential to significantly enhance the speed, accuracy and alignment of public health guideline development and consolidation.

In summary, the ability of AI, particularly LLMs, to understand and process natural language offers a significant opportunity to enhance and potentially transform the way large volumes of evidence are synthesized, a foundational element of clinical and public health decision-making. While the potential applications are wide-ranging, for example, in conducting systematic reviews and consolidating health guidance, current implementation mostly remains at an early stage. Further evaluation is needed to understand both the benefits and the risks, ensuring that the integration of these technologies supports the methodological rigour and transparency required in health care and public health evidence generation.

Should AI be used in evidence synthesis?

While there are several opportunities to leverage the strengths of LLMs to improve the speed and accuracy of evidence synthesis in health, several considerations arise when considering adoption, especially regarding the accuracy and transparency of the results.

System readiness and integration considerations

Accuracy presents a significant challenge when applying AI to evidence synthesis. LLMs are known to generate hallucinations, outputs that appear plausible but are factually incorrect or unsupported by the source data. This poses serious risks when AI is used to extract, summarize or interpret scientific findings. A systematic review of generative AI in evidence synthesis found that when used for searching, AI models missed 68–96% of studies. In the screening of studies, the AI model made erroneous inclusion decisions in 0–29% of cases, and 4–31% of data extractions were incorrect (Clark et al., 2025). While AI tools may improve efficiency, their outputs require careful verification and cannot currently be relied upon in isolation from expert review.

Governance, ethics and rights considerations

Another key limitation in applying AI to evidence synthesis is the issue of transparency. Systematic reviews are built on the principles of methodological rigour, transparency and reproducibility, especially in the development of search strategies and selection criteria. Traditional methods rely on Boolean logic and structured protocols that can be audited, replicated and updated. In contrast, the use of LLMs or other AI tools often involves prompt-based querying, which lacks standardization and is inherently less transparent. This black box nature of LLMs can make it difficult to fully understand how studies were identified or excluded, potentially undermining confidence in the findings. These limitations can be partially mitigated by integrating XAI techniques, which aim to make AI decisions more interpretable, and by applying LLMs in a way that aligns with established methodological standards and reporting frameworks (NICE, 2025).

Frameworks designed to guide users in interacting with generative AI through prompt engineering can also enhance the robustness of AI use (Lo, 2023).

In conclusion, while AI holds considerable promise in accelerating evidence synthesis and expanding analytical capacity, its limitations in transparency and accuracy must be carefully managed. Ensuring appropriate human oversight, adopting best practice standards and continually evaluating these tools in real-world contexts will be essential to safely and effectively integrating AI into evidence generation for health.

2.7 Case studies of AI in health in Europe

This section presents a series of case studies illustrating how European health systems are harnessing AI to transform public health and health care delivery. From the bold reform of primary care in the Autonomous Community of Catalonia, Spain, to Finland's national SOTE-AI ecosystem, these examples showcase diverse strategies for integrating AI into clinical workflows, data infrastructure and policy frameworks. Each case highlights technological innovation and the importance of ethical alignment, professional engagement and system-wide coordination. Whether through the use of ambient scribes, predictive diagnostics or secure data or processing environments, these initiatives reflect a shared commitment to using AI as a tool for equity, efficiency and resilience. By examining real-world implementations across Europe, including Germany, Slovenia and the United Kingdom, this section offers valuable insights into the opportunities and challenges of scaling AI responsibly in public health. Together, these stories provide a roadmap for policy-makers, practitioners and researchers seeking to navigate the evolving landscape of intelligent health systems.

2.7.1 Finland: creating an AI ecosystem and transforming public services with AI

Strategic coordination, ethical governance, infrastructure development and capacity-building for AI in health and social care

Policy-makers face many difficult choices in seeking to maximize the benefits of AI in health systems and minimizing the risks. Finland provides an example of a plan to achieve this balance with its vision for the SOTE-AI ecosystem, which is a national network of health and social welfare actors that was launched in 2024 by the Ministry of Social Affairs and Health. This ecosystem brings together over 250 organizations, ranging from public authorities and research institutions to private companies, with a shared mission: to responsibly harness the power of AI to enhance health and social services nationwide for patients and

the community. The ecosystem is not a rigid structure but a dynamic, voluntary network. It thrives on collaboration, shared learning and experimentation. Key national actors support its coordination, and, together, they are shaping a future where AI is not just a technological tool but a strategic enabler of better care, more efficient services and empowered professionals.

A cornerstone of this plan is a series of 10 pilot projects designed to test AI solutions in real-world settings with the aim of gaining valuable insights into their effectiveness and scalability and, also, the legal implications. With €2.3 million in funding provided by the Ministry and the Well-Being Services Counties (self-governing regions organizing health care and social welfare services), these pilot projects are expected to pave the way for broader national adoption.

Finland's vision goes beyond experimentation. In spring 2025, the ecosystem undertook a vision-building initiative to define a shared national direction for AI in the social affairs and health sectors. This process will result in a white paper that outlines ethical principles, strategic goals and policy recommendations for leveraging AI. The vision also includes a comprehensive assessment of AI's potential impacts, drawing on data from both the ecosystem's pilot projects and other national initiatives. To support this vision, the ecosystem is investing in shared infrastructure and capabilities. Studies are underway to explore the feasibility of a national AI infrastructure that could streamline collaboration and reduce costs. Other initiatives include the development of a Finnish-language AI model tailored to the SOTE context, a national library of reusable AI models and the creation of synthetic datasets and test environments to facilitate the safe and effective development of AI.

Recognizing that technology alone is not enough, the ecosystem is also focused on building human capacity. In 2025, it will produce national teaching materials for educational institutions, helping to equip future professionals with the skills needed to work with AI. It also supports ongoing learning through events such as "AI mornings" and encourages organizations to share their experiences and best practices. A key enabler of this collaborative spirit is the Well-being Regions AI Network. This group meets every three weeks to share updates, identify common challenges and foster partnerships among its members. It plays a crucial role in maintaining a real-time overview of AI development across the country and ensuring that knowledge flows freely between regions.

The ecosystem also addresses the risks and regulatory challenges associated with AI. A dedicated risk management group is identifying potential hazards, evaluating mitigation tools and preparing a report to guide future efforts. Another group is tackling the complex intersection of AI and the MDR, aiming to clarify legal interpretations and support developers in navigating certification processes.

These efforts are essential to ensuring that innovation does not come at the expense of safety or compliance, both of which are foundational to building and sustaining public trust.

On the legislative front, the Ministry is preparing new laws to enable the use of AI in areas such as predictive health care and social welfare. Amendments to the Client Data Act are also being considered to accommodate AI applications better. These legal reforms are critical to removing barriers and creating a supportive environment for AI innovation.

Finland is positioning itself as a global player in digital health. In 2025, the ecosystem will develop a national AI offering and marketing strategy to promote Finnish solutions internationally. It will support organizations to participate in EU projects and global events, and establish a structured approach to managing international business opportunities. To measure its progress, those who created the ecosystem have set clear targets: growing its membership to 250 organizations, increasing the number of newsletter subscribers to 1500, launching at least one joint project and organizing four ecosystem events. They are also aiming to secure a leadership plan for 2026, ensuring the ecosystem's sustainability beyond 2025.

In summary, Finland's 2025 roadmap for the SOTE-AI ecosystem is a comprehensive and forward-looking strategy. The roadmap combines practical experimentation with the creation of a strategic vision, infrastructure development with skills training and national coordination, coupled with proactive international outreach and seeking collaboration with others working in this rapidly changing field. It reflects a deep commitment to using AI not just as a tool, but as a transformative force for better, fairer and more efficient social and health care services.

Practical applications of AI to improve productivity, equity, early intervention and safety in public service delivery

Finland is undertaking a series of ambitious AI projects aimed at transforming its social and health care services. These initiatives have diverse goals and collectively reflect a national commitment to leveraging AI to enhance efficiency, equity and quality of care.

A recurring theme across the projects is the drive to improve productivity and reduce the administrative burden on professionals. For instance, the automatic clinical documentation project, piloted with the Gosta Aide tool at a local health centre in 2024, demonstrated that AI could significantly reduce the time physicians spend on paperwork. Building on this success, the automatic clinical documentation project plans to expand to other professional groups such as nurses and physiotherapists, with a phased national rollout beginning in 2025. In another example, the AI assistant for professionals is being developed to support

health care workers by streamlining needs assessments and service guidance. This assistant is expected to improve work productivity and enhance the quality of care by enabling more time for direct client interaction.

Another major focus of the AI projects is on early intervention and risk prediction. The AI-assisted assessment of child service needs is a response to the issue of child welfare notifications having doubled over the past decade, resulting in an overwhelming burden in many regions of Finland. This project uses AI to compile assessment summaries from large datasets and identify risk factors that threaten child welfare, enabling professionals to meet statutory deadlines and provide timely support. Likewise, the AI-based prediction of functional capacity changes is designed to anticipate declines in people's physical or cognitive abilities. By analysing assessment data, the system enables proactive interventions, potentially preventing costly procedures and improving outcomes for older people, the unemployed or rehabilitating clients.

Access and equity are also central concerns of the AI projects. Two projects are piloting real-time AI-based interpretation tools to overcome language barriers in both health care and social services. The LingAI project, for example, allows clients to communicate with professionals without needing a traditional, human, interpreter, which is especially valuable when interpreters are unavailable. This tool is being tested in housing-related and low-threshold guidance services, where timely communication is critical. In the Finnish context, low-threshold guidance services refer to easily accessible, non-stigmatizing support services that individuals can use without complex procedures, referrals or eligibility requirements. These services are designed to be approachable and user-friendly, particularly for individuals who may be hesitant to seek help. By reducing delays and improving service quality, these tools aim to make public services more inclusive and responsive.

Data integration and utilization underpin many of the initiatives. Considerable effort has gone into compiling background data on users, addressing the challenge of navigating vast amounts of data stored in health care and social service systems. AI can retrieve and summarize historical records in plain language, highlighting key themes such as substance use or self-harm. This enables professionals to make informed decisions quickly, particularly in high-pressure situations such as emergency assessments or initial consultations. Similarly, a cancer imaging project seeks to enhance diagnosis by positron emission tomography (PET) scans. This involves developing a user interface for a model algorithm to detect cancer to enable faster and more accurate diagnoses. It integrates data from histopathological findings, genetic tests and treatment histories to predict treatment responses. This allows personalized treatment plans and optimized

follow-up schedules to be generated, thereby improving both efficiency and patient outcomes.

Some projects are notable for their scalability and potential national impact. A digital obesity treatment initiative, for instance, aims to increase productivity by 50% through the use of AI-assisted nutritional therapy and computer vision-based meal analysis. With a current staff of 14 professionals, the goal is to serve 4500 patients annually, up from 40 to 60 patients per day per professional. Given that obesity-related diseases cost Finland an estimated €5 billion annually, this project represents a scalable solution that offers scope to reduce health care costs significantly while improving patient outcomes.

The development of an AI-powered medication risk tool addresses another critical area: patient safety. In Finland, 13% of health care costs are attributed to correcting medical errors, and nearly a quarter of emergency hospital visits by older people are due to medication-related issues, 90% of which are preventable. This tool identifies high-risk patients in emergency departments and directs them to either a pharmacist or a physician. Its effectiveness will be measured by reductions in adverse events, improved accuracy of medication lists and a reduction in medication changes.

While the projects vary in their technological approaches, from NLP to computer vision and multimodal AI, they share a common goal: to make public services more responsive, efficient and equitable. Some are in the early pilot stages, while others are already being scaled up, reflecting a dynamic and evolving landscape of AI integration in Finland's public sector. Together, these initiatives form a cohesive ecosystem that addresses immediate service delivery challenges and lays the groundwork for a more intelligent, data-driven future in social and health care services. Finland is actively seeking international collaboration in the area of health AI, especially for how health data can be used in AI development.

2.7.2 Germany: a case study from the Centre for Artificial Intelligence in Public Health Research at the Robert Koch Institute where AI is advancing public health research

Established in 2021, the Centre for Artificial Intelligence in Public Health Research at Germany's Robert Koch Institute aims to advance public health research using AI technologies. The Centre combines the Robert Koch Institute's expertise in infectious and noncommunicable diseases, bioinformatics, digital epidemiology and big data analysis with AI and machine learning methods to create a renowned hub for data-driven public health research. The Centre is structured into five research areas.

- **AI fundamentals**: Research focuses on developing innovative AI technologies for public health. This includes designing fundamental AI methods, exploring ethical considerations and ensuring the quality and reliability of AI models applied in health.
- **Phylogenomics**: This area focuses on the intersection of phylogenetics and epidemiology, applying AI approaches to analyse infectious disease transmission dynamics, developing AI-based models of pathogen phylogeny and spread, and deciphering the evolutionary processes shaping modern and ancient pathogens.
- **Image analysis**: The research group focuses on developing and applying algorithms to analyse images, using imaging data, such as microscopy, MRI, CT, X-ray and satellite data. The group creates methods to extract information from biomedical images and adapt existing algorithms for infectious disease research and diagnostics.
- **Climate and societal analysis**: This unit focuses on strategically managing and expanding research at the intersection of climate change, society and public health. It develops AI methods to create digital models and early warning tools to strengthen resilience to climate-related health risks. The unit combines climate modelling, disease modelling, network analysis and computational social and behavioural science to achieve this goal.
- **Visualizations**: Research focuses on developing computational solutions combining visualization and AI. The unit creates methods that use visualization to explain and explore AI models and data analytics, aiming to help users better understand and control these tools. It also investigates ways to address challenges in visualizing data, including uncertainty, accessibility and bias.

Emerging research conducted at the Centre demonstrates the diverse applications of AI in public health:

- **Infectious disease management**: Machine learning has been used to analyse the Lassa virus surface protein and help differentiate Lassa virus lineages, which can support the development and improvement of diagnostics, treatment and outbreak monitoring. Machine learning has also been applied to guide the analysis of a serological multiplex assay, facilitating the distinction between immune responses to mpox infection and those induced by vaccination.

- **Environmental health**: Researchers have developed a multi-scale machine learning model that estimates heat-related mortality in Germany across varying temporal and spatial scales, including regional, annual and projected future risks under different climate change scenarios. These insights can support targeted public health measures during heatwaves and improve long-term health system resilience.

- **Public health communication**: Neural networks and clustering approaches were applied to analyse social media posts about mask-wearing, helping to identify sentiment patterns across different topics. These methods can provide valuable insights for tailoring public health messaging and addressing misinformation.

The Centre offers a three-year PhD programme, which admits new students annually, to train scientists in interdisciplinary research at the intersection of AI and public health.

2.7.3 Slovenia: strategically embracing AI in health care

Slovenia is emerging as a thoughtful and ambitious player in the global race to harness AI, particularly in the health care sector. The country has set out a comprehensive vision for integrating AI into its digital and societal fabric in two key policy documents, which both have health as a key priority: Digital Slovenia 2030 – the strategy for Slovenia's digital transformation by 2030 (Republic of Slovenia, 2023) and the National Programme for AI 2025 (Republic of Slovenia, 2021).

At the heart of Slovenia's digital transformation strategy is the belief that AI should serve people. This human-centric approach is evident in the way the government has framed its ambitions: not merely to adopt AI technologies, but to do so in a way that enhances quality of life, ensures ethical safeguards and promotes inclusive access to innovation.

The Digital Slovenia 2030 strategy sets out the overarching framework and represents the country's comprehensive response to the evolving challenges of digitalization. It outlines a strategic roadmap to guide the country's digital development up to 2030, focusing on key areas such as gigabit infrastructure, digital economy transformation, modern digital public services, the transition to Smart Society 5.0, cybersecurity, digital skills and inclusion, as well as supportive ecosystems for innovation and the green transition.

Building on this foundation, the National Programme for AI 2025 provides a more detailed roadmap. It outlines 10 strategic objectives, including the deployment of AI in public services and the development of reference implementations in

key sectors. Health care is singled out as a domain where Slovenia already has a critical mass of expertise and infrastructure to lead. The programme also includes the development of a national data space as well as pilot environments within health systems.

Slovenia's evolving AI ecosystem includes a dynamic network of public and private institutions. An important partner is the Department of Intelligent Systems at the Jožef Stefan Institute in Ljubljana, which develops and applies advanced AI methods to address societal challenges in health, the environment, energy and beyond. To further strengthen national capabilities, Slovenia has launched a public tender to establish a Competence Centre for Artificial Intelligence, a collaborative consortium of companies, research institutions, academic bodies, innovation clusters and nongovernmental organizations active in the AI domain. The Ministry for Digital Transformation leads initiatives to enhance AI-related skills and competencies, ensures appropriate regulation and builds public trust in AI technologies. On the international stage, Slovenia hosts the International Research Centre on Artificial Intelligence under the auspices of UNESCO, and has joined the D9+ Group of EU digital frontrunners, affirming its commitment to innovation and digital leadership.

AI applications in health – opportunities and barriers

The adoption of AI in Slovenia's health care system is still in its early stages, with few fully established practices. However, several promising initiatives and pilot projects are already underway. Among these, there are successful examples of leveraging AI to increase efficiency, improve access to care and support early diagnosis and treatment. For example, a hospital implemented the WoShi AI system to automate staff rostering for around 110 employees. Creating optimal schedules manually used to take about seven working days; WoShi reduced this to one day and now handles about 90% of the tasks, greatly decreasing the administrative load (ALGiT, 2023). Predictive algorithms are also being leveraged to help forecast long-term risk of cardiovascular and metabolic conditions among young people, through the SmartCHANGE Horizon Europe project, which is coordinated by the Jožef Stefan Institute (SmartCHANGE, 2025).

In addition to using AI to improve resource allocation and support with early diagnosis, there are also emerging uses of LLMs to provide patient and health care staff support. HomeDoctor is an innovative Slovenian prototype e-health platform that integrates GPT-4o with the Insieme system and verified national medical knowledge, with the aim of delivering 24/7 virtual health care support. Designed for both mobile and desktop use, it combines the capabilities of general LLMs with local health information with the goal of providing accurate, context-aware medical advice. While the platform shows strong potential, it faces challenges

related to data privacy, the accuracy of user input and the need for Slovenian-trained language models. Ongoing development focuses on expanding medical databases, enhancing NLP capabilities, ensuring compliance with national data protection regulations and exploring the use of open-source foundation models that can be deployed locally in a secure processing/data environment. Future directions include the integration of advanced features such as predictive analytics and personalized health recommendations. Ultimately, this work seeks to lay the groundwork for modernizing Slovenia's digital health infrastructure and alleviating pressure on the health care system (Kocuvan et al., 2024).

A recent report on the use of call centres and digital tools in Slovenian primary health care facilities reveals a strong foundation for future AI integration. Current systems used to manage patient communication are built on structured digital platforms and standardized workflows, which AI technologies could effectively build upon. For instance, several facilities have established call centres that handle high volumes of patient calls using digital platforms for call routing and performance tracking. These systems could be enhanced with AI for automated triage, prioritization and sentiment analysis. Additionally, web-based tools already allow patients to request appointments, submit medical documents and communicate with health care staff using standardized forms and templates. AI could further improve these services through chatbots, predictive appointment scheduling and NLP to interpret free-text inputs. In summary, while AI is not yet implemented in the current Slovenian family medicine call handling systems, the existing digital infrastructure and standardized processes provide a strong basis for future integration (Lunežnik, 2024). AI has the potential to significantly enhance efficiency, patient satisfaction and equitable access to care, especially in high-demand environments.

Conclusion

Slovenia is actively embedding AI across the health care sector, with applications ranging from workforce management and digital therapeutics, to predictive diagnostics and national infrastructure. These initiatives demonstrate a strong trend towards combining clinical workflows, patient empowerment and data-driven policies underpinned by AI.

The country's strategic embrace of AI in health care is both ambitious and grounded. By aligning technological innovation with societal needs, ethical principles and international collaboration, it is positioning itself as a leader in human-centric AI. As the initiatives outlined in Digital Slovenia 2030 and the National Programme for AI 2025 continue to unfold, Slovenia offers a compelling example of how a small nation can make a significant impact in the age of intelligent health.

2.7.4 The Autonomous Community of Catalonia, Spain: using AI as a lever for reform to transform primary care services

In Spain, the Autonomous Community of Catalonia is taking bold steps to modernize its health care system and has placed the strengthening of primary care at the heart of this transformation, viewing it as the cornerstone of an effective, equitable and sustainable health system. Within this vision, AI is no longer a future aspiration but an active component of care improvement.

A reform grounded in primary care

The CAIROS health system reform programme, launched in late 2024, prioritizes transforming primary care as the first and most urgent step. While the primary care system in Catalonia has long been admired (Kringos et al., 2013), it is now experiencing decline, resulting in reduced access to family doctors and nurses. Building on a broad consensus formed through extensive expert consultations and a participatory process known as Open CAIROS, the authorities in Catalonia have approved a set of 25 measures organized into 10 lines of action to reform the health system, starting with primary care (Martí & González López-Valcárcel, 2025). This transformation is being piloted in 27 selected primary care centres, now called Integral Health Centres of Reference, which represent a diverse geographical and organizational sample of Catalonia's 377 primary care teams. These Integral Health Centres of Reference are responsible for testing new models of organization, professional roles, financial incentives and technology adoption, including AI.

Introducing AI-based clinical support

In April 2025, Àxia Clinical Support, an AI-powered chatbot, was introduced across the pilot 27 Integral Health Centres of Reference after a positive evaluation in a research environment (Fuster-Casanovas et al., 2025). Two months later, Àxia Clinical Support was scaled up to all primary care teams in the region. This generative AI system, which uses a structured and hierarchical RAG process, provides diagnostic and therapeutic guidance tailored to the Catalan context. Crucially, the guidance given adheres strictly to the local clinical guidelines and protocols employed by Àxia. It is available to family doctors, paediatricians and primary care nurses through the electronic health record system. Rather than replacing clinical judgement, this AI agent acts as a digital copilot, assisting decision-making and enhancing the clinical process. It is designed to support professionals while respecting their professional autonomy and clinical choice.

After six months of deployment, the chatbot had processed 46 748 threads and 183 359 messages. User satisfaction is high (89%), adoption rate low (21%)

and adherence is quite high (85%). These figures suggest that while the chatbot is valued by the clinicians who use it, managers face challenges as they seek to increase uptake, calling for better communication and training. Although this was not originally anticipated, this system makes it possible to exploit qualitative metrics, such as the types of questions asked, and tailor training provision to the needs of diverse populations.

Streamlining documentation with ambient scribes

Five ambient AI scribes from different vendors are currently being tested in parallel and evaluated in real settings. These tools capture spoken clinical interactions between health professionals and patients and convert them into structured clinical notes. The resulting documentation, organized into the reasons for the consultation, examinations, medical history and procedures, is then ready for direct integration into electronic health record. These AI scribes allow professionals to re-focus on patient interaction, improving the quality of the clinical encounter and reducing the administrative burden that often contributes to burnout by reducing the time they spend at their computers. Provisional monitoring metrics confirm high patient satisfaction (94%) and widespread professional acceptance (78%), although this is higher among nurses (84%) than doctors (72%).

Ambient scribes are also being piloted in other Spanish regions under the coordination of the Ministry of Health of Spain, providing a broader framework for comparison and facilitating national scalability. The pilot ran until the end of September 2025, and followed a standardized methodology designed to assess compliance with the core technical and functional requirements that had been defined for their potential use in the Spanish National Health System.

The objective of the pilot was to identify the desirable technical and functional features to deploy across the Spanish National Health System, such as multilingual capability, required performance, clinical coding functionality and ease of use, and to produce a report enabling the development of common model specifications for subsequent procurement and implementation by regional health services. The pilot involved 44 primary care centres and a representative group of family physicians, paediatricians and nurses. By the end of September 2025, more than 5263 consultations had been carried out, with the satisfaction level of the solution rated at 4.4 out of 5 for the patients, and 4.2 out of 5 for the professionals. In the future, an internal marketplace is planned for the Spanish National Health System to serve as a repository of validated algorithms and a hub for system-wide early demand identification, complemented by a centralized procurement mechanism to facilitate product acquisition by the regional public health services.

Towards a multi-agent AI ecosystem

These two AI-based applications are part of a broader strategy to incorporate intelligent digital agents across the care continuum, before, during and after the clinical visit. The vision for the use of AI is not limited to isolated tools, but rather to building a multi-agent ecosystem that enhances the capacity and experience of both professionals and patients. It is a step towards rethinking workflows and redefining productivity in primary care.

Lessons learned from implementing AI in primary care

The introduction of AI-based clinical support and ambient scribes across Catalonia's primary care teams has provided valuable lessons that can help shape future efforts to scale AI in health care in a safe, meaningful and sustainable way.

- **Clinical alignment is non-negotiable.** Tools that generate diagnostic and therapeutic suggestions must be aligned with local clinical protocols and guidelines to be accepted by professionals. The credibility and safety of the clinical support tool relied on its strict adherence to official protocols in force in Catalonia, which fostered early adoption.

- **Trust is built through transparency and co-design.** Professionals were more likely to embrace AI when they understood how the tool worked and had a voice in its configuration. Early involvement of clinicians in the selection, piloting and feedback process proved essential for building trust and usability.

- **Ambient scribes relieve cognitive and administrative overload, but require adaptation.** AI transcription tools significantly reduced time spent on documentation, improving professional well-being and patient interaction. However, adaptation to different consultation styles and specialities required iterative refinement. One-size-fits-all models did not perform equally well in all settings.

- **Integration with existing systems is key to impact.** For both tools, seamless integration into the electronic health record will be critical. Fragmented digital environments or cumbersome workflows undermined the perceived value of the tools and created friction in adoption.

- **Training and change management cannot be an afterthought.** Even intuitive tools require time for professionals to adapt. Structured onboarding, peer learning and dedicated support teams helped smooth the learning curve and avoided resistance rooted in uncertainty or misinformation.

- **Continuous evaluation fuels improvement and scale-up.** Ongoing monitoring of usage, professional satisfaction and clinical outcomes allowed for real-time adjustments and supported a data-driven decision to scale the tools beyond the initial pilot projects.
- **AI should augment, not replace, professional judgement.** Clinicians responded positively when AI was presented as a supportive agent not a substitute, and reinforcing the message that AI assists (but does not decide) was key to overcoming scepticism and ensuring ethical alignment.
- **Equity requires proactive design choices.** Implementing AI in primary care must consider varying levels of digital infrastructure and professional digital literacy. Tailoring deployment strategies to different team capacities and regional contexts helped reduce the risk of widening gaps.

Conclusion

Catalonia's AI strategy for primary care is not about automating health care. It is about augmenting it. By embedding AI into the daily routines of primary care, the CAIROS reform aims to deliver better care, more accessible services and a more rewarding professional experience. If successful, this approach could offer a replicable model for other health systems seeking to harness digital transformation for the common good.

2.7.5 United Kingdom: using the OneLondon Secure Data Environment to create a secure, scalable and AI-ready data infrastructure for health care

Overview

The OneLondon Secure Data Environment (SDE) represents one of the most advanced and integrated efforts in the United Kingdom to create a secure, scalable and AI-ready data infrastructure for health care. Originating from the Discover-NOW initiative launched in 2019, the OneLondon SDE was developed in partnership with Health Data Research UK (HDR-UK) and the London Health Data Strategy (NHS Digital, 2025). It evolved through the collaboration of London's five Integrated Care Systems and three Health Innovation Networks. Its development has been underpinned by a £3.6 million investment from NHS England as part of the broader Data Saves Lives programme, which seeks to embed regional SDEs across the country.

Infrastructure and vertical integration

The OneLondon SDE is a federated, cloud-based platform providing access to linked, longitudinal health data for over 2.8 million Londoners, with plans to scale to more than 10 million by 2026. It integrates pseudo-anonymized primary care records, commissioned acute care datasets and legacy electronic health records. To support interoperability, reusability and consistent analytics, these datasets are structured using the Observational Medical Outcomes Partnership (OMOP) common data model. The OMOP provides a shared vocabulary and data structure, allowing analytic tools and AI models to be developed once and applied consistently across multiple datasets and care settings. This approach promotes open code sharing, reproducibility and efficient deployment of AI pipelines.

The SDE brings together three interconnected components. The London Data Service, hosted in north-east London, handles data extraction, linkage and provisioning of pan-London primary and secondary care datasets, supporting secure analytics for both NHS operational use and research. The London Research and Analytics Environment, originating from the HDR-UK Discover-NOW Hub in north-west London, governs secure research environments for academic and commercial users and provides a unified pan-London analytics platform. The AI Centre for Value Based Health Care acts as a national centre of excellence, providing expertise and technology to enable multimodal data integration, federated analytics and the deployment and validation of machine learning models.

The technical infrastructure underpinning the SDE supports the full AI life-cycle: model development, validation, deployment and post-market monitoring. This vertical integration ensures that AI models can be aligned with clinical workflows, embedded into real-world settings and evaluated against long-term health outcomes, all within the governance framework of the NHS.

Real-world applications and capabilities

The OneLondon SDE is currently being used to deliver real-time population health insights that inform decision-making across the NHS. Applications include optimizing care pathways for long-term conditions, identifying high-risk patients for targeted intervention and highlighting opportunities for proactive prevention and treatment optimization. Built on reusable and adaptable infrastructure, the SDE provides transferable analytic building blocks that can be adapted for multiple use cases and care settings, demonstrating how AI-driven insights can be translated into practical, actionable improvements in both patient care and population health management.

The AI Centre for Value Based Health Care has been extending the capabilities of the SDE to include multimodal data. Its Federated Learning and Interoperability Platform enables structured health records in Observational Medical Outcomes Partnership format, multimodal imaging data and imaging metadata to be analysed and used for AI model training across hospital trusts without physically transferring sensitive data. This preserves the security and privacy of patient data. The AI Centre for Value Based Health Care is also standardizing data from proprietary hospital systems and using CogStack, an advanced NLP platform, to transform clinical notes and other narrative text into structured, analysable data. This ensures that both primary and secondary care data, including pathway and previously unseen data, can be surfaced into the SDE ecosystem, supporting richer analyses for precision medicine and population health management.

Patient and public engagement

To ensure public trust and accountability, OneLondon has engaged in a wide range of deliberative public engagement initiatives. These include hosting the first-ever Citizen's Summit, which explored expectations and complex issues around the use of health and care data. Patient and public outreach is also conducted through Integrated Care Boards, patient committees in AI Centre NHS Trusts, and the Discover-NOW programme. This engagement helps ensure that data use is aligned with societal expectations, ethical principles and the public interest.

Policy relevance

By facilitating robust evaluation, reproducibility and public trust, the OneLondon SDE sets a benchmark for health data environments. Its architecture offers the building blocks for policy-makers seeking to develop AI-ready infrastructure that meets the demands of modern health care delivery and research. It embodies a layered model, from secure data storage and governance to real-time analytics and continuous monitoring, that aligns with both national and EU policy directions. The platform highlights how policy levers such as funding, regulation, procurement and public engagement can be mobilized to build sustainable and responsible AI ecosystems in health care.

2.7.6 Conclusion

These case studies underscore a shared European ambition to embed AI meaningfully into health systems, not as a replacement for human expertise, but as a strategic enabler of better care, improved efficiency and greater equity. Whether through clinical support tools in Catalonia, national infrastructure in Finland or SDEs in London, successful implementation hinges on ethical

alignment, professional trust and system-wide integration. The importance of co-design, transparency and continuous evaluation emerges as a consistent theme, alongside the need for robust data governance and inclusive digital strategies. Together, these examples offer a compelling vision of AI as a public health ally, one that augments human capabilities, strengthens resilience and supports more responsive and sustainable health care systems. As countries move from pilot projects to full-scale implementation, these lessons provide a valuable foundation for shaping future policy and practice in the age of intelligent health.

Chapter 3

Questions facing health policy-makers making decisions on AI

KEY MESSAGES

System readiness and strategic integration

- Evaluating and safely implementing AI in health requires more than technical performance. It demands robust validation, ethical and economic assessment and alignment with real-world clinical needs, supported by frameworks such as HTA and resilient, interoperable infrastructure.
- AI can improve efficiency and reduce costs, but its benefits depend on thoughtful integration, including managing risks such as data drift and system failures, ensuring transparency and preparing the health workforce for shifting roles and responsibilities.
- As AI becomes a strategic asset in global health policy, sustainability and international cooperation are increasingly vital, with environmental impacts and regulatory divergence prompting the need for shared standards that uphold human rights and public trust.

Governance, ethics and human rights considerations

- Ethical AI in health care requires a multi-layered approach, combining strong regulation, technical safeguards and human oversight, while also addressing deeper questions about embedding human values in increasingly autonomous systems.
- Accountability and transparency are essential, supported by clear legal frameworks, robust oversight mechanisms and safeguards to prevent unintended or unethical behaviour, especially in high-risk areas such as health care.
- Privacy, fairness and inclusivity must be built into AI systems through responsible data practices and inclusive design, with the goal of protecting civil liberties and promoting equitable access to AI benefits.
- Sustainable and democratic AI development depends on secure infrastructure, environmental responsibility and meaningful participation from diverse stakeholders, ensuring that AI serves the public interest and aligns with shared societal values.

Building on the foundational questions explored in Chapter 1 and the review of AI applications covered in Chapter 2, this chapter examines the practical challenges that policy-makers and health professionals may face as they navigate evolving technological and regulatory landscapes. As outlined in section 1.2, AI systems span a broad spectrum, from early foundational models based on statistics and rules, to traditional machine learning approaches that identify patterns in data, through to deep learning systems capable of handling complex data inputs and, finally, to modern, more autonomous and generative AI systems. Each stage in this evolution brings distinct opportunities and risks. For example, early statistical models are relatively transparent and interpretable but may oversimplify reality; machine learning approaches can uncover patterns in large datasets but may amplify biases; deep learning systems can support complex decision-making but often lack transparency; and modern generative or agentic AI systems offer powerful capabilities but introduce challenges around trustworthiness, hallucinations and accountability. This diversity of AI underlines why a one-size-fits-all view of AI is insufficient: different systems pose very different challenges depending on their design and context of use. Bearing this in mind is essential when engaging with the questions presented here, which are organized into two thematic areas: system readiness and strategic integration, and governance, ethics and rights. While most of these questions are not specific to any single geography, the discussion is framed in reference to the EU AI Act and its implications for Member States. Table 3.1 (page 121) provides a high-level summary of the policy goals addressed in each area, how these align with the provisions of the EU AI Act and highlights additional considerations and examples of good practice identified through our research and engagement across European countries.

Table 3.1 *Summary of the AI in health policy goals addressed in this chapter, how these align with the provisions of the EU AI Act, additional considerations and examples of good practice*

Policy goal	Relevant provisions in the EU AI Act	Scope and reach of the EU AI Act	Related policies or regulations (selected)	Other considerations and sources of good practice
Establish robust, adaptable regulatory frameworks that support innovation while promoting human-centric, trustworthy and ethical AI	Entire Act and risk-based classification	As a Regulation, it is directly enforceable across the EU, so Member States cannot introduce stricter or divergent national rules on AI systems that fall within the scope of the AI Act, unless the Act allows for such flexibility. The exceptions are areas not fully harmonized by the AI Act (e.g. national security, certain public sector uses). Member States may still adopt complementary rules, provided they do not conflict with EU law. Member States may also enforce stricter rules under other EU laws, such as GDPR. The EU AI Act came into force in 2024, but its rollout is taking place in phases until full application in August 2027	Council of Europe Framework Convention on Artificial Intelligence Data protection regulation (e.g. GDPR) Medical device regulations (e.g. EU MDR and IVDR) ISO standards (e.g. ISO 42001)	Regulating AI systems is complex, particularly for generative and agentic AI, and deployers must understand the legal requirements. Several factors, including values and principles, regulation and policy, technical design, governance frameworks and the human–machine interaction, influence ethical development and use of AI
Ensure accountability, safety and human oversight	Article 5 (Prohibited AI practices), Article 14 (Human Oversight), Article 15 (Accuracy, Robustness, and Cybersecurity), Article 16–27 (Obligations of Providers and Deployers), Article 50 (Transparency Obligations), Article 72 (Post-Market Monitoring), Chapter V (general-purpose AI models), Chapter XII (Penalties)	Certain AI practices are prohibited. The Act imposes specific requirements for high-risk AI systems for effective human oversight and quality management systems (including accountability frameworks). Transparency obligations are defined for providers and deployers of certain AI systems. Member States must establish rules for imposing penalties and enforcement measures, and the Commission can fine developers of general-purpose AI models if they break the rules	Data protection regulation (e.g. GDPR) Medical device regulations (e.g. EU MDR and IVDR) ISO standards	Accountability in AI is complex and has different dimensions, including compliance, reporting, oversight and enforcement. Level of human oversight required will depend on the type of task and associated risk. Explainable AI can improve transparency, but it has limitations when used with advanced AI models, and decisions on the degree of human oversight needed must consider the complexities of human–machine collaboration, such as automation bias

> > *continues*

Policy goal	Relevant provisions in the EU AI Act	Scope and reach of the EU AI Act	Related policies or regulations (selected)	Other considerations and sources of good practice
Promote fairness, equity and inclusion	Article 5 (Prohibited AI practices), Article 10 (Data and Data Governance), Recitals and voluntary codes of conduct (Article 95)	Certain AI practices are prohibited. The Act requires bias mitigation practices, especially with high-risk AI systems. Recitals and voluntary codes of conduct promote these principles, but are not binding obligations in themselves. Recitals are, however, highly persuasive in interpreting the binding Articles of the Regulation and courts and regulators draw on them to clarify intent behind provisions, resolve textual ambiguities and understand context and objectives of the law. Codes of conduct may be used as evidence of good faith or due diligence in compliance assessments or legal disputes and serve as benchmarks for ethical AI development, especially in sectors where formal regulation is still evolving	Council of Europe Framework Convention on Artificial Intelligence International human rights instruments	It is important to consider the impact of AI on fairness, equity and inclusion across the life-cycle of the technology. For example, certain generative AI training approaches can be unfair, and the adoption of AI can be inequitable due to gaps in digital exclusion. Similarly, measures to promote fair, equitable and inclusive AI should be considered across the AI life-cycle. For example, bias in AI can emerge at different stages, such as the algorithm design or the training data used
Invest in and promote digital and AI literacy and workforce capacity	Article 4 (AI literacy), Recital 20, voluntary codes of conduct (Article 95)	All providers and deployers of all AI systems should ensure AI literacy among staff. High-risk systems must be developed in a way that human overseers can effectively understand the capabilities and limitations of the AI system (e.g. automation bias)	Data protection regulation (e.g. GDPR)	Efforts to increase AI literacy among health professionals should include an understanding of the capabilities and limitations of various AI systems, the complexities inherent in human–machine interactions, and the ethical considerations associated with their deployment. AI literacy among the patients and the general population is essential to ensure effective and equitable access to public-facing technologies, as well as meaningful participation in decision-making processes

> > *continues*

Policy goal	Relevant provisions in the EU AI Act	Scope and reach of the EU AI Act	Related policies or regulations (selected)	Other considerations and sources of good practice
Ensure data protection and privacy, and build secure, interoperable AI infrastructure	Article 10 (Data and Data Governance), Article 15 (Accuracy, Robustness and Cybersecurity), Article 57 (AI Regulatory Sandboxes), Recitals	Requires specific data governance considerations for high-risk AI systems. Each Member State must create at least one AI Regulatory Sandbox (controlled environments where AI systems can be developed, tested, and validated under the supervision of national authorities before being placed on the market)	Data protection requirements are set out in the GDPR, which is complemented by the EU AI Act and the European Health Data Space Regulation	There can be different data access arrangements to train and test models, such as secure processing/ data environments and federated learning approaches, while maintaining future-proof infrastructure that supports the development and deployment of AI. This includes interoperable infrastructure, integrated into health pathways and enables the reusability of data, models and analytical pipelines
Support participatory and decentralized AI development	Recitals and voluntary codes of conduct (Article 95)	EU AI Act encourages diverse and inclusive AI design and development	Council of Europe Framework Convention on Artificial Intelligence	Governments must consider ways to promote participatory approaches through public involvement and promote decentralized AI development through strategic investment
Integrate environmental and socioeconomic sustainability	Recitals and voluntary codes of conduct (Article 95)	Encourages assessment and minimization of the impact of AI systems on environmental sustainability	European Green Deal	Consider and measure the environmental impact of AI across its value chain, including data centres, with investment in efficient and sustainable AI innovation
Foster global cooperation	Recitals	International cooperation encouraged by the EU AI Act	Council of Europe Framework Convention on Artificial Intelligence	Federated approaches can enable the secure development of AI models across different geographies

AI: artificial intelligence; EU: European Union; GDPR: General Data Protection Regulation; IVDR: In Vitro Diagnostic Medical Devices Regulation; MDR: Medical Device Regulation.

3.1 System readiness and strategic integration

This set of questions focuses on the practical and strategic challenges of integrating AI into health systems and public institutions. It includes considerations around AI performance and clinical effectiveness, real-world value, safety, sustainable digital infrastructure, cost-effectiveness, workforce implications, geopolitical dynamics and environmental impact. These questions are intended to help policy-makers think through the operational, technical and long-term implications of AI development and adoption, including how to align innovation with broader system goals and constraints.

3.1.1 How do we evaluate the technical performance and clinical effectiveness of AI models in health?

As noted in section 2.1, two fundamental questions to be answered when applying AI in health care are whether the technology can reliably perform the intended task and what impact it has on health outcomes. This section asks how AI tools in health can be evaluated for technical performance, with a focus on supervised machine learning models and generative models, and how clinical effectiveness can be assessed. While technical performance is initially assessed during development under controlled conditions, it also requires continuous monitoring after deployment to ensure that performance remains stable in real-world settings. By contrast, clinical effectiveness, measuring actual health outcomes, is almost always evaluated during deployment through approaches such as before-and-after studies, intervention studies or randomized controlled trials.

Technical performance

Different AI tools require different evaluation approaches. In health care, we can broadly distinguish between predictive models (for diagnosis, triage and prognosis) and generative models (that produce text, such as summaries or chatbot responses). Each requires its own set of metrics and validation standards.

We first consider the evaluation of supervised models with a ground truth. This is a type of machine learning where the model is trained on a labelled dataset, that is, a dataset where the correct answers (the ground truth) are already known. The ground truth might be, for example, a diagnosis made from a medical image confirmed by a panel of radiologists. These models must be tested on data they have not seen during training, using clinically relevant end-points (such as a diagnosis or event). Typically, a held-out test set from the same source is used for internal validation, but this alone does not guarantee generalizability. At a minimum, developers should demonstrate strong performance on both internal validation datasets, drawn from the same population as the training data but

not used during model training, as well as on external validation datasets that are entirely separate and independent from all development data (training, validation and internal test sets) (Wiens et al., 2019). This distinction between internal and external validation is critical because internal performance often overestimates real-world effectiveness, which is why regulatory frameworks and HTA increasingly emphasize external evidence.

However, many evaluations fall short (Oddy et al., 2024; Siontis et al., 2021). Test datasets may lack real-world representativeness or can be drawn from overly controlled environments that may be too similar to training data. In diagnostic models, performance can be inflated by testing models on a non-complex spectrum of cases (Tseng et al., 2021). For example, an AI model tested to detect lung cancer in patients with either late-stage cancer, large nodules or no cancer may perform extremely well, but be significantly poorer in a real-world setting where medical imaging signs may be subtle and scattered with unrelated incidental findings. In labelled data, the gold standard labels themselves may be flawed or outdated, particularly when generated through retrospective coding or non-blinded review. Further, external validation performance may not generalize to all deployment settings and thus should be interpreted in context.

As with any model, AI-based or not, model performance should be reported with statistical confidence, and sample sizes should be sufficient to detect whether statistical significance is clinically meaningful, particularly across relevant subgroups such as age, ethnicity or deprivation. This relates back to the question of equity, which is addressed separately in this chapter but is critical in evaluating whether tools work fairly across the population or if there is potential for a tool to exacerbate disparities.

Evaluation also demands transparent and comparable performance metrics. Precision, recall, F1 score and area under the receiver operating characteristic (AUC-ROC) curve remain the core indicators for classification models (Kelly et al., 2019) (Box 3.1, page 126). Each captures different aspects of performance, and health professionals appraising AI evaluations should be cautious if any single metric is offered as definitive. For regression-based models, such as predicting a value (systolic blood pressure, for example), ground truth is still required, but performance is typically evaluated based on how well the model explains the variance in the data, using metrics such as the R-squared score. Additionally, various error measures are used, including mean squared error, mean absolute error and root mean squared error.

Beyond supervised machine learning models, it is important to note that unsupervised models are not validated against predefined labels, but evaluated for coherence, internal consistency and clinical relevance (Chen et al., 2020;

Wiens et al., 2019). For example, clustering algorithms may help stratify patients with long COVID, but cannot be directly judged by accuracy. Their value lies in insight generation, not decision-making, and they must be interpreted with caution (Rajkomar et al., 2019).

Box 3.1 *Measures of performance of prediction by AI*

- **Precision** measures the proportion of positive predictions that are actually correct. In other words, if a model identifies 100 patients as having sepsis, and only 80 truly do, the precision is 80%. This metric is especially important when the cost of false positives (incorrectly identifying someone as ill) is high.
- **Recall**, which is analogous to sensitivity, measures the proportion of actual positive cases that the model successfully identifies. If 100 patients truly have sepsis, and the model correctly identifies 70 of them, the recall is 70%. This is crucial in health care settings where missing a diagnosis can have serious consequences.
- The **F1 score** is the harmonic mean of precision and recall. It provides a single metric that balances both concerns, especially when there is an uneven class distribution (e.g. a significantly higher number of sepsis cases than non-sepsis cases). A high F1 score indicates that the model is performing well in both identifying true cases and minimizing false alarms.
- The **AUC-ROC curve** plots the true positive rate (recall) against the false positive rate at various threshold settings. The area under this curve (ranging from 0 to 1) gives an aggregate measure of performance across all classification thresholds. A model with an AUC close to 1 is considered very good at distinguishing between classes. For example, a model predicting sepsis in hospital patients might show high accuracy, the overall percentage of correct predictions, but this can be misleading if sepsis is rare. A model could achieve high accuracy simply by predicting that no one has sepsis yet fail to identify any actual cases. In such scenarios, precision and recall offer a more informative picture. A model with high precision but low recall might correctly identify a few sepsis cases while missing many others, potentially putting patients at risk. Conversely, a model with high recall but low precision might catch most sepsis cases but also generate many false alarms, overwhelming clinicians with unnecessary alerts and reducing trust in the system (Saito & Rehmsmeier, 2015).
- **R-squared** is the proportion of variance in a model outcome that is explained by the model rather than chance. It is a method used to assess the accuracy of regression-based models. An R-squared of 0, therefore, means the model is no better than guessing the mean value. An R-squared of 1 indicates that the model explains all the variability in the outcome, i.e. accurately models the true relationship. A negative R-squared can occur when the model performs worse than guessing.

> > *continues*

Box 3.1 *continued*

- **Mean squared error** is a measure of accuracy in regression-based models. It is the average of the squared differences between predicted and actual values. It penalizes larger errors more than smaller ones, making it useful when large prediction errors have a high clinical impact.
- **Mean absolute error**, another regression-based accuracy measure, is the average of the absolute differences between predicted and actual values. It is easier to interpret than mean squared error and less sensitive to large errors, making it useful when outliers should have less influence.
- **Root mean squared error** takes the square root of the mean squared error. Again, it is used in regression-based models and is very interpretable due to its units being the same as the original outcome (Hodson, 2022).

We next consider the evaluation of text-generating (generative) AI tools. These are systems that use machine learning, especially LLMs, to produce human-like text based on a given input or prompt. These tools can generate anything from simple sentences to complex documents, depending on their design and training.

LLMs present new challenges for performance evaluation. Due to their nuanced outputs, performance must be assessed multi-dimensionally to understand linguistic fluency, clinical relevance, safety and factual accuracy. Both automated and human steps are essential to contribute different insights. Automated metrics allow speed, consistency and scalability, particularly in arduous tasks such as comparing words or phrases overall in a model output compared to a ground truth reference text. Tools such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BLEU (Bilingual Evaluation Understudy) are widely used in machine translation or summarization to evaluate accuracy (Box 3.2, page 128). While these tools detect structural similarity and surface-level errors, they fall short in their ability to identify nuanced discrepancies, detect omissions of critical findings and distinguish between incorrect conclusions, such as diagnoses or advice (Nguyen et al., 2024).

To address this, semantic similarity metrics like the BERTScore have been introduced (Zhang et al., 2020). These use transformer-based models to evaluate the meaning of an LLM output relative to a reference. However, these still struggle with clinical nuance, which in health care can have significant clinical implications. Due to these limitations, human evaluation remains indispensable in health care contexts and is reflected by the human-in-the-loop emphasis in many regulations, such as the EU AI Act and the MDR.

Domain experts, such as clinicians, patients and public health professionals, are uniquely positioned to assess semantic accuracy, completeness, trustworthiness, tone and safety. Structured evaluation frameworks, such as QUEST, provide a systematic way to assess model output from a user-centric and evidence-based perspective (Tam et al., 2024). Human reviews are increasingly being harnessed to apply clinical judgement to detect subtle but critical problems that automated metrics miss, including hallucinations (Asgari et al., 2025).

As with prediction models, the outputs of LLMs can be validated both internally, typically by the companies developing them, and externally. In this context, external validation often involves testing the model in real-world simulations or pilot studies. It is essential that these scenarios accurately reflect the complexity and variability of real clinical environments to ensure that the AI system's performance is realistically assessed.

Finally, newer advances in LLM-as-a-verifier are beginning to blur the line between generator and evaluator. LLM-as-a-verifier or judge refers to LLMs that assess their own models or outputs. For instance, Amazon Bedrock has introduced LLM-as-a-judge tools that evaluate systems on dimensions, such as correctness, helpfulness, response refusal and harmfulness, providing normalized scores and natural language explanations. Moreover, such verifiers hold particular potential for RAG systems. As discussed in section 1.2, RAG-based models rely on external documents to provide context. LLM-as-verifier techniques have been applied to ensure that the generated output remains grounded in the retrieved evidence, improving the faithfulness, relevancy and trustworthiness of responses. RAG triad, a conceptual framework for evaluating RAG systems, uses three key metrics: context relevance, groundedness and answer relevance. Metric-driven evaluation tools, such as RAGAS, DeepEval and VERA, use LLM-based evaluation to detect hallucinations and assess generator accuracy relative to context, helping optimize and validate RAG pipelines systematically (Es et al., 2025). These methods are still experimental and not yet robust enough for clinical deployment, but they show promise.

Box 3.2 *Measures of performance of LLMs*

- **BLEU (Bilingual Evaluation Understudy)** is an automated metric that compares how closely an AI-generated text matches a reference by checking for overlaps in words or short phrases. It is widely used in tasks like translation and summarization for measures of completeness. While useful for spotting surface-level errors, BLEU does not capture deeper meaning or clinical appropriateness.

> > *continues*

Box 3.2 *continued*

- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)** is another automated metric, commonly used to assess summaries. It measures how much of the reference text is captured in the generated output. ROUGE focuses on recall, which is helpful when completeness is important, but, like BLEU, it does not evaluate factual accuracy or clinical safety.
- **BERTScore** uses a language model to assess how similar the meaning of the generated text is to a reference. It goes beyond word-matching and tries to evaluate the semantic content of a response. This is a useful step forward, but BERTScore still struggles with complex clinical reasoning and may not detect serious errors in advice or interpretation.
- **QUEST** is a structured framework for evaluating the quality of AI-generated health content from a human-centred perspective. It looks at aspects such as clarity, relevance, trustworthiness and evidence use. QUEST is especially valuable in clinical settings, where expert judgement is needed to assess tone, nuance and safety of responses.
- **PEMAT (Patient Education Materials Assessment Tool)** is designed to assess how understandable and actionable patient information is. It is often used by health communicators to evaluate whether materials are suitable for the general public. For LLM-generated outputs aimed at patients, PEMAT helps determine if advice is not only correct but also easy to follow.
- **DISCERN** is a tool originally developed to evaluate the quality of written health information. It uses a checklist-based approach to assess accuracy, balance and reliability. Although designed before modern AI, DISCERN remains a useful framework to assess whether LLMs are producing safe and trustworthy consumer-facing health content.

Clinical effectiveness

Even a highly accurate model may be clinically ineffective if it cannot be integrated into care safely and meaningfully. Integration into data and pipeline infrastructures are discussed in section 3.1.4; here, we aim to address how to determine if AI system integration into clinical workflows improves outcomes, reduces harms or supports better decision-making when compared to standard practice (Cruz Rivera et al., 2020).

Evidence of effectiveness usually emerges from empirical studies, such as before-after studies, controlled intervention studies, or the gold standard randomized controlled intervention studies (Box 3.3, page 130). All of these allow us to see how well the tool performs not just on unseen data, but also when embedded in clinical practice, as these tools are not immune to pressures of user trust, usability in workflows and real-world unforeseen events. As with any study looking at changes over time, to capture the more longer-term impacts, the length of study and clinical relevance of outcomes assessed must be critically analysed.

Importantly, the comparator matters. Too often, AI tools are evaluated against poor-quality or outdated care, which inflates claims of effectiveness. Instead, tools should be measured against current best practice, whether that is expert clinician judgement, validated risk scores or structured pathways. For example, a triage tool for stroke must demonstrate an advantage over existing protocols, not merely outperform no triage at all (Topol, 2019).

To ensure rigorous design and transparent reporting of clinical evaluations of AI systems, guidelines such as SPIRIT-AI and CONSORT-AI have been developed (Cruz Rivera et al., 2020; Liu et al., 2020). These frameworks extend existing clinical trial standards to address the unique challenges posed by AI-based interventions, including defining dynamic algorithm behaviour, human and AI interaction and real-world deployment context. Using such standards helps ensure that trials evaluating AI tools, especially randomized controlled studies, produce valid, interpretable and reproducible evidence of clinical effectiveness.

Box 3.3 *Measures of clinical effectiveness*

- **Before–after studies** measure outcomes in a health care setting before and after introducing an AI tool. They are simple and quick to implement and help detect whether care improved following implementation. However, they are vulnerable to other changes happening at the same time (such as staffing or policy shifts), which can make it difficult to isolate the AI tool's true impact.
- **Controlled intervention studies** compare outcomes between two groups, one using the AI tool and another continuing standard care, without random assignment. These studies help account for broader contextual factors and can offer stronger evidence than before-after designs, but they still carry risks of bias, as the groups may differ in important ways.
- **Randomized controlled trials** are the gold standard for evaluating whether an AI tool is effective. Participants or clinical settings are randomly assigned to use the AI tool or to continue with usual care. This ensures the groups are comparable and that any differences in outcomes are more likely to be due to the AI tool itself. Randomized controlled trials offer the highest level of evidence but can be complex, costly and time-consuming to run.

3.1.2 How do we assess the real-world value of the AI system?

Effective evaluation of AI systems in health requires more than technical benchmarking and clinical effectiveness in controlled environments. This section provides a brief overview of the application of established HTA methods to the use of AI in routine practice, including issues of implementation.

As noted in section 2.1, HTAs are widely used internationally to inform decision-making about the adoption and use of new health technologies. Rather than focusing solely on clinical effectiveness or cost, HTAs take a multidimensional approach that encompasses a broad set of considerations relevant to patients, clinicians, policy-makers and health systems. In Europe, a widely adopted framework is the EUnetHTA Core Model, developed by the European Network for Health Technology Assessment. This model identifies nine core domains essential to comprehensive evaluation: 1) the health problem and current use of the technology; 2) technical and functional characteristics; 3) safety; 4) clinical effectiveness; 5) cost and economic implications; 6) ethical aspects; 7) organizational impact; 8) social implications; and 9) legal considerations (Kristensen et al., 2017). These domains ensure that evaluations are systematic, transparent and relevant to real-world decision-making.

Although HTA is a robust, widely used method for assessing new technologies, recent analyses suggest it may require adaptation for AI-based tools. A study conducted within the EUnetHTA initiative highlighted that, while its Core Model covers essential domains, expert evaluations revealed that many critical facets of AI assessment, such as algorithmic bias, human oversight, transparency and trustworthiness, were often inadequately addressed or entirely missing from current evaluations. As such, there is increasing recognition that AI-specific criteria must be integrated into existing HTA frameworks to ensure evaluations remain comprehensive and contextually relevant (Di Bidino et al., 2024). In January 2025, the European Commission announced an update to the EU rules on HTA, which sees the potential for selected high-risk medical devices being assessed at EU level to avoid duplication of HTA and a joint consensus across Member States (European Commission, 2025b). In parallel, there are also studies looking at how AI might help improve the HTA framework, particularly in the predicting cost-effectiveness domain (Ramezani et al., 2025).

For any new medical device, including those incorporating AI, regulatory approval by the relevant authority (such as the Medicines and Health Care Products Regulatory Agency in the United Kingdom or a Notified Body in the EU) is mandatory before market entry and HTAs. Regulatory approval typically confirms compliance with safety and performance requirements under frameworks such

as the EU MDR or the United Kingdom MDR. However, these approvals often rely on clinical investigations focused on short-term performance and safety rather than long-term real-world effectiveness.

Both the EU MDR and the United Kingdom MDR require clinical evaluation and, where necessary, premarket clinical investigations following Good Clinical Practice standards (ISO 14155) (ISO, 2020a). Yet, these studies are controlled and rarely replicate real-world complexity. Post-market surveillance is therefore a legal obligation under the MDR and is reinforced by ISO/TR 20416:2020 guidance (ISO, 2020b), requiring manufacturers to continuously collect and analyse data throughout the device life-cycle, including through post-market clinical follow-up.

While the EU MDR and the United Kingdom MDR integrate post-market surveillance as a legal obligation, in practice, post-market surveillance evidence is limited at the point of market entry. Post-market surveillance plans focus on future data collection, but this evidence typically matures over time, meaning that the initial approval is still based primarily on controlled studies rather than real-world performance. Consequently, HTA bodies often require additional real-world evidence to evaluate long-term effectiveness, comparative performance and cost-effectiveness before recommending widespread adoption. For AI-enabled technologies, this challenge is amplified due to factors like algorithm updates, data drift and workflow integration, underscoring the need for early planning for real-world evidence generation beyond regulatory requirements (Chouffani El Fassi et al., 2024).

It is important to recognize that the evaluation of AI systems must continue beyond deployment. These systems are susceptible to data drift, where changes in patient populations, clinical practices or health care environments may lead to a decline in performance over time (Guan et al., 2025). To ensure continued safety and effectiveness in real-world settings, ongoing monitoring, recalibration and revalidation are essential (Efthimiou et al., 2024). To support this, developers should build post-deployment safeguards into the system from the outset. These include mechanisms such as audit trails, risk logs and version control, which not only facilitate monitoring but also enhance accountability, debugging and regulatory compliance (Brojka et al., 2024; Mokander et al., 2022). Although the infrastructure needed for continuous monitoring and updates is discussed later in this chapter, technical readiness must begin during model development, not just at the point of deployment.

However, as previously noted, regulatory frameworks often lag behind the rapid pace of AI innovation, particularly in the case of adaptive or continuously

learning systems. In practice, this makes it more challenging to obtain approval for models that evolve. As a result, some organizations may choose to deploy static models, which are easier to regulate but may be less adaptable to future changes in clinical practice or patient populations.

Finally, evaluation must reflect the realities of clinical care, where a clear-cut ground truth is not always available. Unlike tasks with binary outcomes or objectively measurable labels, much of medicine operates in a grey area. Clinical decisions often fall within a spectrum of reasonable judgements, shaped by patient preferences, contextual factors and the professional experience of clinicians. AI systems designed for these complex environments must be able to accommodate this inherent uncertainty. Rather than aiming to replace human reasoning, such tools should be designed to support clinical judgement, offering suggestions or highlighting considerations without asserting a single "correct" answer. Treating AI as an infallible source of truth risks oversimplifying complex cases and undermining the nuanced, context-sensitive thinking that high-quality care requires. This is especially relevant in areas where no definitive ground truth exists, such as treatment planning, prognostic discussions or triage decisions, where multiple valid approaches may coexist. Therefore, evaluation frameworks must be flexible enough to account for reasonable variability and avoid penalizing outputs that differ from reference texts but remain clinically appropriate.

In summary, while evaluating the technical performance of AI tools is essential, effective assessment in health care must go beyond metrics alone. Measures such as precision, accuracy and hallucination rate are important for understanding how well a model performs its core tasks, but they represent only one part of a much broader evaluation framework.

To ensure AI tools are not only accurate but also usable, safe and beneficial in real-world health care settings, performance evaluation must be integrated with wider clinical, operational and societal considerations. The HTA frameworks introduced earlier offer a valuable foundation for this integration. However, they must be adapted to meet the unique demands of rapidly evolving AI technologies.

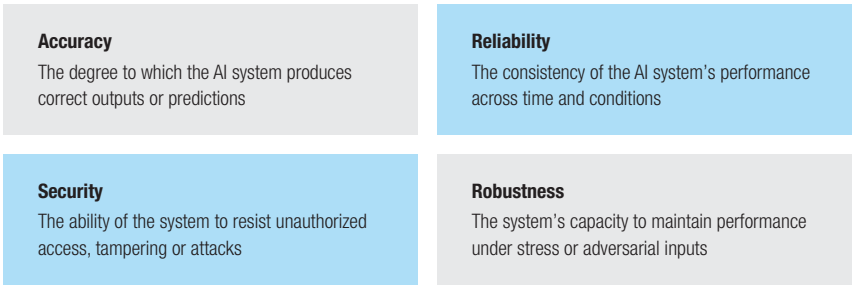
As health care continues to evolve, so too must the frameworks used to assess new technologies, including AI. Traditional HTA domains, such as safety, economic impact, ethical considerations and organizational effects, remain critical. Yet, they must now be expanded to include emerging concerns such as algorithmic bias, dynamic model updating, explainability, trustworthiness and system-level interactions. These additions are essential to ensure that AI tools are not only technically sound but also aligned with the values and complexities of modern health care.

3.1.3 How can we make AI safe?

Implementing measures to ensure the safety of AI systems has emerged as a key priority for policy-makers. Unlike traditional software, advanced AI systems can exhibit unpredictable behaviour, reflecting their complexity, potential for inaccuracy and ability to hallucinate. When AI systems fail or are exploited by malicious actors, the consequences can be severe and, in some cases, life-threatening. Consequently, a failure to embed safety into AI throughout its life-cycle is not an option. Achieving this requires a life-cycle approach, addressing safety during development and maintaining it throughout deployment.

One way to define safety is the AI system's ability to function reliably without causing unintended harm. AI systems can produce erroneous outputs, malfunction under stress, misinterpret unusual inputs or be deliberately manipulated, raising risks for individuals, businesses and public infrastructure. From a technical perspective, safety by design, meaning that safety considerations are integrated from the earliest stages of AI development, is a part of the EU MDR and the United Kingdom MDR. This includes accuracy, reliability, security and robustness (Fig. 3.1). Accuracy can be operationalized in several ways, including the fraction of correct outputs or the percentage of errors that occur. Reliability is a measure of consistency. Security refers to a system's capability to protect its architecture and information from unauthorized access or damage. Finally, robustness refers to the ability of the AI system to maintain performance even when faced with adversarial inputs, unexpected situations or partial system failures (Leslie, 2019).

Fig. 3.1 *Aspects of the safety of AI tools*



Source: Adapted from Leslie (2019).

There is a broad range of risks that can threaten these safety dimensions. Some are unintentional, related to weaknesses in model design, others can be intentional and pose significant threats to cybersecurity. There is a wide range of adversarial techniques and corresponding defence mechanisms, and the nature and severity of these risks vary depending on the type of AI model involved (Pelekis et al., 2025). In essence, adversarial threats can be defined as deliberate

perturbations to model inputs designed to induce incorrect outputs (Goodfellow et al., 2015). For example, an adversarial attack may aim to extract information about the training data or the model architecture, potentially leading to the leakage of confidential or sensitive information. Another type of attack involves data poisoning, whereby an adversary manipulates the model's training data or parameters to degrade its performance. This degradation can, in turn, enable future exploitation, such as by injecting malicious data into the training dataset, creating vulnerabilities that can be leveraged over time (Pelekis et al., 2025). If health data are compromised, the consequences can be severe, including identity theft and misuse of medical records. To mitigate this, there is a provision for EU Member States to develop AI regulatory sandboxes enabling AI systems to be tested in controlled environments before deployment to identify and mitigate critical safety risks (European Union, 2024). Even with robust design measures, AI systems face new challenges once deployed in real-world environments, requiring ongoing monitoring and governance.

Safety in deployment refers to the systems and safeguards in place to ensure AI systems remain safe, effective and trustworthy once integrated into real-world settings. This can look like continuous monitoring, alerts and risk detection, which are necessary due to the potential for machine learning models to experience a decline in performance over time due to a phenomenon known as data drift (Moreno-Torres et al., 2012). This arises because many AI models are trained on historical data and their underlying parameters remain fixed after deployment. This can create problems when they are used in dynamic environments where the characteristics of incoming data evolve, for example, as a result of changes in the patient population demographics in the case of a clinical AI application. These shifts in the characteristics of the data being processed mean that, over time, it may no longer accurately represent the data on which it was trained. The resulting mismatch can adversely affect the model's accuracy, reliability and overall utility, as well as the safety of decisions based on its outputs.

In addition, one of the challenges for policy-makers is that AI failures may not be easily detectable or explainable. A system may work well during testing but fail catastrophically in real-world conditions. This is especially problematic in autonomous systems, such as self-driving cars or medical diagnostic tools, where errors can lead to injuries, deaths or systemic failures.

To address these issues, policy-makers must enforce rigorous safety standards across the AI life-cycle. Regulatory frameworks address both pre-deployment safety testing and post-deployment monitoring to ensure AI systems remain safe throughout their life-cycle. This includes requiring pre-deployment testing under realistic and diverse conditions, continuous monitoring during operation and the ability to shut down or update systems if risks emerge rapidly. The EU

AI Act incorporates a risk-based classification of AI systems, in which certain AI applications are prohibited, and higher-risk applications face stricter oversight and compliance obligations.

The EU AI Act complements other medical device regulations, such as the EU MDR and the In Vitro Diagnostic Medical Devices Regulation (IVDR) (Aboy et al., 2024). In addition, organizations should ensure that AI tools are formally registered with the relevant regulatory authorities. In England, for example, primary care practices are beginning to adopt ambient scribe software. However, organizations such as the British Medical Association are calling for caution, highlighting potential risks related to information governance and patient safety. They advise that products should be registered with the Medicines and Health Care Products Regulatory Agency and be compliant with NHS standards before adoption (BMA, 2025). This example underscores the complexities that regulatory systems encounter amid the rapid adoption of emerging technologies and highlights the potential consequences these developments can pose to public safety.

Ultimately, both safety by design and ongoing safety in deployment are essential in ensuring AI safety protects the well-being of the population, the integrity of institutions and public trust in technology. Policy-makers must lead in establishing a governance model that prioritizes caution, transparency and adaptability, so that as AI grows more powerful, it remains aligned with human safety and societal values.

3.1.4 What infrastructure is required to build sustainable AI solutions in the health system?

The sustainable and safe deployment of AI in health care requires a well-structured technical ecosystem that spans multiple elements, including data, storage, computing power and security. This environment will need to support not only the safe development of AI tools but also their integration, monitoring and continuous improvement. This section will focus on the technical infrastructure necessary to support responsible AI throughout its life-cycle.

Open and closed environments are an important distinction to make during AI development and deployment. Open environments allow broader access to AI systems and may use publicly available data or incorporate usage data back into model training. In contrast, closed environments restrict access and prevent raw data from leaving the secure infrastructure. In the context of sensitive health data, particularly under European data privacy regulations such as the GDPR, AI development and deployment should occur in closed environments to ensure compliance and protect patient confidentiality. These secure, closed

infrastructures can include cloud-based SDEs, local on-premises servers with firewall protection or other tightly controlled IT systems where sensitive data are prevented from leaving the environment. This distinction has direct implications for infrastructure design, governance and the technical features required to support safe and responsible AI in health care.

First, a foundational requirement of a sustainable and secure AI ecosystem in health is high-quality, representative and interoperable health data. This includes structured data (e.g. laboratory results, prescriptions) and unstructured data (e.g. text in clinical notes or radiology reports) from various sources, such as primary care, hospitals and population health systems (Rajkomar et al., 2019). Interoperable data requires standardized and well-governed datasets, with mechanisms to securely integrate information across institutions, legacy systems and regions, enabling consistent and meaningful use for AI development and analytics.

Technical infrastructure must also include secure and scalable computing and storage capacity. Developing large-scale models, such as those used for image classification or generative tasks like summarizing discharge notes, requires access to high-performance computing in secure environments that meet standards for privacy, anonymization and cybersecurity (European Commission, 2023b). Computing and storage resources may be co-located or separate, with trade-offs in cost, scalability, speed and integration; careful planning ensures efficient AI workflows while maintaining data security. Many cloud-based infrastructures offer integrated secure storage, computing power and familiar data science tools within tightly governed environments that ensure auditability and traceability of all actions (Sendak et al., 2020).

Once deployed, AI tools will not remain static. Their performance can degrade over time due to model drift, as described in section 3.1.3 (Wiens et al., 2019). Addressing this requires operational infrastructure such as machine learning operations (MLOps). MLOps support versioning, incident logging, model retraining and bias detection (Kreuzberger et al., 2022). These capabilities are especially important in health care, where outdated or misaligned models can lead to clinical harm. For instance, an AI triage tool developed on data from one region may underperform when deployed in another without proper calibration or adaptation, leading to missed diagnoses or misclassified patients (Sendak et al., 2020).

Equally important is the ability to integrate AI tools into clinical workflows. This includes potentially embedding decision-support models directly into electronic health records, ensuring that outputs are presented in clinically meaningful ways and at the right moment in the care process. This capability may require

more complex infrastructure; however, would enable real-time data exchange, model interpretability and feedback from health professionals that can help AI to support, rather than disrupt, care delivery (Shortliffe & Sepulveda, 2018; Topol, 2019).

A consideration that may be possible for some areas within health care is the scope for vertical integration of AI infrastructure, particularly within a closed environment. This refers to environments where data storage or acquisition, model training, validation, deployment and monitoring can occur within a single, coordinated system with standardized pipelines and governance. A benefit of a vertically integrated system is the formation of reusable assets and data pipelines. This means code or models developed in one secure environment can be applied in another secure environment, improving efficiency of data processing tasks and encouraging collaborative working while maintaining strict data access controls (NHS Digital, 2025; One London, 2025). Such integration supports safer, more scalable AI by allowing for real-world performance evaluation, continuous updates in response to data or workflow changes and transparent version control. Implementing a fully vertically integrated system can be complex and resource intensive, particularly for hospitals with legacy IT systems, requiring careful planning and investment.

It is important to be aware of the distinctions between AI tools developed within public health systems and those acquired from commercial vendors. Public sector development, when done well, offers greater alignment with population needs, adherence to open science principles, and prioritization of local clinical interests, while also making it easier to ensure that data will remain within trusted institutions and is governed in accordance with public values (Morley et al., 2020). Retaining data in-house, in a closed environment, can make it easier to protect privacy, provide greater contextual understanding and facilitate the development of models that are more responsive to local epidemiology, clinical practice and language use. By contrast, outsourcing to third parties can introduce dependencies, long-term costs and risks around data ownership, privacy and explainability. However, keeping the data in-house also requires significant resources, particularly in steps such as data labelling for supervised training. Many commercial models are developed in open or semi-open environments, meaning that training data and usage data may not remain fully private. While commercial tools may offer speed of implementation, they often rely on proprietary data and opaque methods and may not generalize well to different populations or health systems. Off-the-shelf models may underperform in local settings, particularly where predictive performance depends on region-specific determinants, underrepresented subpopulations or linguistic variation (USAID Center for Innovation and Impact et al., 2019). Further considerations on model building strategies are provided in Box 3.4 (page 139).

Box 3.4 *Model scope and building strategy*

In certain AI health applications, choices have to be made between developing bespoke models from scratch or adapting existing foundation models. Each approach carries distinct trade-offs in terms of cost, performance, interpretability, customization, resources, bias and domain alignment. These considerations are also different depending on the type of AI model.

A key question facing many health organizations or institutions is whether to train and build AI models in-house or whether to procure off-the-shelf AI models from third-party vendors. Developing custom models can be complex and resource intensive but allows for precise tailoring to specific needs in terms of data, organizational workflows and population characteristics. This can be especially important in domains with unique data modalities or underrepresented populations. Customized models can be optimized for interpretability and regulatory compliance, which can be critical in clinical settings. However, this approach is resource intensive, requiring substantial computational infrastructure, secure processing/data environments, workforce expertise and high-quality data (One London, 2025). The level of resources needed and their complexity will depend also on the type of AI model and the desired application or task. While some institutions may be able to develop their own shallow machine learning models, for example for predicting risk of hospital readmission based on their existing patient data, this approach is almost impossible when it comes to foundation models. In some cases, particularly where real-time or offline processing is needed, institutions may opt for smaller or distilled models that can be deployed on edge devices (IBM, 2025b). These models can be more computationally efficient and suitable for environments with limited connectivity or infrastructure. As noted in Box 1.4 Techniques to improve the performance of LLMs in a specialized task, adapting existing foundation models through transfer learning approaches offers a more scalable alternative (IBM, 2024c). These models, pretrained on vast and diverse datasets, can be fine-tuned using different techniques. However, fine-tuning models can also be an extremely complex task and these models may carry biases from their original training data (Lu et al., 2025; Schramowski et al., 2022). Additionally, their black box nature can hinder explainability and trust in high-stakes environments. A key consideration in fine-tuning models is whether to use open-weight foundation models, which offer greater transparency and customization potential, or closed models, which often benefit from large-scale proprietary training but restrict access and adaptability. This decision can significantly influence trust, explainability and the long-term sustainability of AI solutions. In addition to the knowledge domain, it is important to consider that most of these LLMs have been trained predominantly using text in the English language, which can be detrimental for model performance in other languages. The EuroLLM project aims to bridge this gap by creating open-source LLMs tailored to European languages (EuroLLM team et al., 2024).

Ultimately, the choice between custom development or adaptation hinges on the context and type of application, data availability, the resources available and regulatory requirements.

In summary, sustainable AI systems require a robust and secure technical ecosystem that supports the full life-cycle of AI tools, from development to deployment and ongoing monitoring. This begins with access to high-quality, representative health data from diverse sources, including both structured and unstructured formats. To ensure privacy and regulatory compliance, especially with sensitive health data, these systems are typically implemented in closed, secure environments such as SDEs, where raw patient data never leaves the infrastructure.

Federated platforms also support sustainable infrastructure by allowing models to be trained and updated across distributed datasets without centralizing sensitive data, reducing duplication of resources, lowering storage and computing demands, and enabling collaborative development across institutions. These contrast with open environments, where models may use publicly available data or incorporate usage data back into training.

Equally important is the integration of AI processes, data acquisition, model training, validation and deployment, within a unified system. This allows for real-world performance evaluation, continuous updates and transparent version control. During deployment, models must be maintained through operational tools, like MLOps, which support versioning, retraining, and bias detection, ensuring that models remain accurate, safe and aligned with real-world contexts. AI tools must be embedded into clinical workflows in ways that enhance, rather than disrupt, care delivery. This includes integration with electronic health records and real-time data exchange, ensuring outputs are clinically meaningful and timely. To further improve the quality and diversity of training datasets, governments and institutions can establish secure systems and formal agreements that enable cross-border collaboration while maintaining strict privacy and governance standards.

Public sector development of AI tools offers advantages in aligning with population needs and maintaining data governance, while commercial solutions may introduce risks around privacy, explainability and generalizability. At the same time, private sector innovation can accelerate development and bring technical expertise that complements public efforts. Infrastructure should support both approaches, with mechanisms for local validation and adaptation to ensure safety and equity. Section 2.7.5 provides an example of a London-based effort to develop a secure, AI-ready health care data infrastructure.

3.1.5 Will AI help reduce health care costs?

Reducing the cost of health care remains an urgent priority as systems face rising pressures from ageing populations, an increasing burden of chronic disease and persistent workforce shortages. In this context, AI is often viewed as a promising tool to enhance efficiency and alleviate strain on services. AI applications in health, including clinical decision support, imaging analysis and administrative automation, have the potential to streamline processes and free-up the time of health care staff for patient care (European Commission, 2025a). Estimates from the USA suggest that the "wider adoption of AI could lead to savings of 5 to 10% in United States health care spending, roughly US$ 200 billion to US$ 360 billion annually in 2019 dollars" (Sahni et al., 2023). In the United Kingdom, Lord Darzi's review of health and care in 2018 estimated that greater use of automation and AI in the English NHS could increase productivity by £12.5 billion per year (Darzi et al., 2018).

The financial benefits of AI are not, however, guaranteed and depend heavily on the context in which these tools are developed and implemented. While pilot studies and industry-led evaluations suggest efficiency gains, the evidence base on the economic evaluations of AI in routine health service delivery remains limited (Kastrup et al., 2024). High-quality, independent economic evaluations are scarce, and there is a lack of research on the long-term economic implications of AI in health (Al Meslamani, 2023).

It is important to recognize that the full costs of AI development and adoption of AI tools can be underestimated. These may include not only the initial investment in data infrastructure and computational capacity, but also the ongoing costs of maintaining and updating models, staff training and digital literacy, energy consumption, ensuring cybersecurity and meeting regulatory requirements. Even some large private corporations are postponing several generative AI initiatives due to concerns about the cost of computing (IBM, 2024a). Importantly, these cost implications of AI will vary significantly depending on whether the models are developed and trained in-house, adapted through transfer learning approaches, or purchased as off-the-shelf solutions. The feasibility of each option will depend on the available resources and the type of AI technology required. As an example, an individual hospital might feasibly build a traditional machine learning model using its own patient data, whereas training an LLM from scratch can require investments of millions of euros or dollars (Maslej et al., 2025).

A further challenge is that problems during implementation may offset any potential savings. Health systems must invest in training to ensure that staff can use AI tools safely and effectively. There is also a risk that AI could unintentionally exacerbate inefficiencies or inequalities, particularly if AI models are trained

on biased or incomplete data. In such cases, rather than reducing costs, poorly designed or misapplied AI solutions could lead to inappropriate care, duplication of services or further administrative burdens.

While AI offers exciting possibilities, it is important to recognize that many of the structural pressures facing health care systems cannot be resolved through technology alone. Many gains in health outcomes, and potentially reductions in cost, are likely to come from investments in the wider determinants of health, such as housing, education and social care, alongside robust prevention strategies (Wood et al., 2024). Without addressing these foundational drivers of ill health, prioritizing technological innovations alone risk offering only marginal improvements within systems already stretched to their limits.

In summary, AI has the potential to support cost containment in health care, particularly through improved operational efficiency. However, its role should be seen as complementary to, rather than a substitute for, broader investments in and reforms to the health system. Realizing the economic benefits of AI will require careful evaluation and implementation of the technology.

3.1.6 What will AI mean for the health workforce?

AI has the potential to transform the future of the workforce, creating new opportunities and challenges that require careful consideration. As AI technologies continue to evolve and integrate into a broad range of industries, one of the most pressing concerns for policy-makers is the socioeconomic disruption they may cause, particularly through job displacement. AI can potentially boost productivity and efficiency significantly, but it can also automate tasks traditionally performed by humans, creating a major shift in the labour market that requires proactive policy responses. In health care, the integration of AI also raises fundamental questions about the role and scope of health care professions, particularly regarding responsibilities and accountability in patient care and decisions that can directly impact the health of individuals and populations.

Planning and preparing for the future health workforce requires an understanding of how AI may change the nature of work through the support or replacement of tasks, as well as by creating entirely new roles. An analysis by The Health Foundation, in the United Kingdom, outlines a framework for navigating the impact of technology on the health care workforce. In line with other analyses, this framework proposes breaking down health roles at the task level and distinguishing between tasks that strengthen, supersede, support or substitute for human capabilities. These four categories are determined by two key dimensions: whether the technology matches or exceeds human performance, and whether

it primarily assists or replaces humans in carrying out the task (Moulds & Horton, 2023).

Given the nature of much of health care, AI-driven automation is likely to have less impact than in other sectors (Filippi et al., 2023). This is driven by several factors, including the complexity of certain competencies in health care, the human interaction component of patient care and the limited number of health care jobs which consist of fully automatable tasks (Moulds & Horton, 2023). In addition, evidence shows that even if the demand for job postings in the health care industry requiring AI skills is growing, it is doing so at a slower pace than other industries (pwc, 2025). The integration of these technologies raises important questions regarding both the intended objectives and the practical implications of their use. In particular, it is important to clarify whether the primary goal of these technologies is to improve the quality of care, enhance productivity or reduce operational costs and how the time saved through automation of routine tasks will ultimately be utilized. While proponents argue that such efficiencies will enable health professionals to devote more attention to direct patient care, others caution that achieving these practical benefits may come at the cost of eroding patient–doctor relationships and dehumanizing medicine (Akingbola et al., 2024).

In addition, navigating clinical or public health decision-making supported by a machine will also challenge health professionals, defined by their specialized knowledge, rigorous training and self-regulation. The abilities of certain forms of AI, particularly LLMs and agentic AI, in the future, are reshaping their roles by challenging traditional scopes of practice. This can take two complementary forms. One is adopting AI solutions to replace health professionals, for example, with automated image analysis in radiology or cytopathology. The other is the employment of less-skilled and, consequently, cheaper health workers, whose lesser ability is augmented by AI tools (McKee & Correia, 2025). These developments are potentially problematic. As described in earlier chapters, AI tools, while powerful, are not without limitations. They can perpetuate biases embedded in their training data, fail to generalize across populations or produce false confirmations of incorrect diagnoses. In addition, this shift risks undermining the essence of professional work. Professionals are not merely technicians; their expertise lies in synthesizing knowledge, navigating ethical dilemmas and building trust with those they serve. Delegating these complex responsibilities to machines or less-qualified workers risks eroding public confidence in professions.

Policy-makers and professional bodies face the daunting task of striking a balance between innovation and tradition. They must ensure that AI is integrated responsibly, with robust regulations to safeguard quality and accountability. This

includes defining roles for emerging technologies and non-traditional workers, providing comprehensive and regularly updated training, and maintaining stringent ethical standards. They must also take a long-term perspective. Overreliance on AI tools can erode professional expertise. If health professionals become overly dependent on AI, they may lose critical diagnostic and decision-making skills. A study exploring the neural and behavioural consequences of using LLMs to support essay writing found that while the AI tool provided immediate help, this benefit came at a cognitive cost. Participants who relied on LLMs demonstrated lower performance compared to those who did not, as measured by neural activity, linguistic complexity and behavioural indicators. These findings raise important concerns about the potential long-term consequences of overreliance on LLMs for complex cognitive tasks (Kosmyna et al., 2025).

Unlocking the potential of these transformative technologies will also require investing in education and training systems that prepare workers for the future. This includes promoting science, technology, engineering and mathematics education, digital literacy and, just as importantly, lifelong learning programmes that allow adults to reskill and adapt as job requirements change. It will be important to recognize the need for safety nets to support displaced workers, although this is beyond the scope of health policy-makers. Unemployment benefits, income support and social services must be flexible enough to cover gig and freelance workers, many of whom are likely to be most affected by automation.

The future of professions in the age of AI will hinge on the ability of health decision-makers and professionals to adapt without compromising their core values. As described in Chapters 1 and 2, this requires a nuanced understanding of both the opportunities and the limitations presented by emerging AI technologies, as well as the ability to critically evaluate which lower-risk tasks can be safely automated and which will continue to necessitate human oversight, judgement, trust and expertise. In health, and particularly in medicine, it remains unlikely that a substantial proportion of tasks will be fully automated in the near future (Topol, 2019).

3.1.7 How might AI systems influence geopolitical dynamics?

Governments increasingly recognize the potential of AI to further many of their broader goals, from increasing economic growth to enhancing their military capabilities and expanding their international influence. Although these considerations may appear remote to many health policy-makers, their scope of action will inevitably be shaped by the intense global competition among nations striving to dominate AI research, infrastructure, talent and deployment. Without deliberate international cooperation, this competition risks devolving

into an AI arms race, eroding trust and heightening the likelihood of misuse (Schmid et al., 2025).

Policy-makers must grapple with the dual challenge of advancing national AI capabilities while promoting global norms and ethical standards. Many governments have identified a strategic need to invest in AI for both civilian and military applications (Clapp, 2025). Such investments may include achieving control over strategic resources such as high-quality datasets and cloud infrastructure. In this context, AI emerges not only as a tool of economic progress but also as a lever of geopolitical power. This environment creates serious risks, with important implications for health. In the absence of robust international norms, countries may feel compelled to deploy untested or unsafe AI systems to gain a strategic advantage. Such dynamics could lead to unintended escalations, the proliferation of autonomous weapons, and human rights violations (Meacham, 2023). These challenges underscore the need for coordinated international frameworks to ensure that AI development and deployment align with shared values and collective security, as exemplified by the Council of Europe Framework Convention on Artificial Intelligence (Council of Europe, 2024) and WHO's *Ethics and governance of artificial intelligence for health* (WHO, 2021).

In addition, technological breakthroughs such as the LLMs released by DeepSeek, a China-based company, demonstrate how advances can reduce development costs and computational requirements, enabling smaller players to compete with dominant firms. This development has disrupted the technology market within China and internationally, paving the way for more diverse AI ecosystems (Conroy & Mallapaty, 2025). Other examples include the availability of LLMs, such as LLaMa 2, that are free of charge for research and commercial use, which have accelerated the democratization of the generative AI landscape.

As discussed in section 1.4, different regions are taking markedly divergent regulatory approaches. These frameworks will strongly influence how AI is developed and deployed in each country, thereby shaping their competitive positioning in the global AI landscape. National and international regulatory frameworks and shared AI principles, such as those developed by the Organisation for Economic Co-operation and Development, are critical to protecting fundamental rights as the technology evolves (OECD.AI Policy Observatory, 2025). However, there is a real risk that developers and investors may concentrate their efforts in jurisdictions with looser regulations, potentially leaving regional groupings such as the EU increasingly reliant on imported AI systems from the USA or elsewhere.

Regardless of how AI development evolves, it is clear that national and international standards and regulation will play defining roles, and that new

global dependencies and complex supply chains will emerge, much as they did with oil or advanced semiconductors. Policy-makers should therefore work towards establishing multilateral agreements that clearly define the rules for using AI, underpinned by a solid human rights framework. Collaborative research initiatives, open science efforts and cross-border talent exchange programmes can further encourage cooperation rather than rivalry. Shared progress in areas such as climate modelling, health innovation and disaster response can help demonstrate AI's potential as a force for the collective good.

Ultimately, AI policy must strike a balance between national interests and global responsibilities. Only through intentional collaboration can governments prevent misuse, manage risks and ensure that the transformative power of AI benefits all of humanity.

3.1.8 Will AI pose a threat to our commitment to the environment?

Policy-makers are now accustomed to considering the environmental consequences of their decisions, such as whether health facilities are energy efficient, whether they can be reached by public transport and whether they employ environmentally friendly methods for disposing of clinical waste. Increasingly, they must also account for the environmental footprint of the AI systems they procure and deploy.

The first issue to consider relates to energy and water consumption and carbon emissions. Training and operating AI models, particularly large-scale deep learning systems, require substantial computational power. Data centres powering the training of these models often consume vast amounts of electricity, a significant portion of which is derived, in many countries, from non-renewable energy sources, leading to increased carbon dioxide emissions. After model training, AI deployment and model updates will require ongoing energy consumption, which is expected to continue increasing as people increasingly incorporate these technologies into their daily lives. A frequently overlooked factor, beyond energy consumption, is the substantial volume of water required to cool data centre infrastructure (Zewe, 2025). Health policy-makers should prioritize partnerships with data providers who utilize renewable energy sources and consider establishing emissions and water use reporting standards for AI vendors in the health care sector.

The second relates to specialized computer hardware, which will further impact the environment through the indirect impact of their manufacturing and transportation, as well as through the generation of electronic waste. The development and deployment of AI systems often require specialized hardware, including high-performance computing components that may have short life-cycles due to rapid technological advancement. The resulting electronic waste can

pose serious environmental and public health risks, particularly in low-income countries where electronic waste is often exported and poorly managed. In addition, as noted in section 1.3.3, AI global supply chains can further reinforce exploitative dynamics involving unsustainable mining or pollution with toxic chemicals, such as mercury and lead (UNEP, 2024). Policy-makers should consider measures that promote the recycling, reuse and responsible sourcing of hardware.

There are many challenges in estimating the global AI energy consumption and its future trajectory. While global energy consumption is projected to rise by 26% by 2030, most of this growth is driven by industries other than data centres. Other estimates from the USA indicate the data centres might be using 4.4% of the country's energy, and this could increase to 7–12% by 2028. However, even if the impact at the global scale seems lower than other industries, the impact that these data centres can have at the local level is large, due to their tendency to concentrate in small geographical spaces (Chen, 2025).

Concerns have also been raised about whether energy demands can be met solely by renewable energy, suggesting that the expansion of AI infrastructure could undermine the limited progress made so far in reducing carbon emissions and addressing climate change. On a planet where many communities already struggle to access clean water, the water usage associated with AI data centres is also projected to continue growing. AIWaterUsage, which started tracking live water consumption of AI infrastructure in April 2025, estimates that OpenAI, the company that released ChatGPT, has used 15 million gallons of water in less than three months across its infrastructure (AIWaterUsage, 2025).

While users should be made aware of the environmental impact of the AI tools they adopt, this challenge cannot be addressed solely through individual responsibility. A collective, coordinated effort is essential to ensure that AI innovation progresses sustainably. Policy-makers should actively promote research and development aimed at improving the energy and resource efficiency of model training and deployment. This includes exploring the use of smaller, more specialized models tailored to specific tasks, rather than relying exclusively on general-purpose LLMs.

The guidance in WHO's *Ethics and governance of artificial intelligence for health* includes promoting sustainable AI solutions (WHO, 2021). The EU AI Act does refer to environmental protection but, while binding environmental impact assessments would have been required in the text approved by the European Parliament, this was not included in the final text, something that environmental campaigners have described as a missed opportunity. Instead, assessment of environmental impacts is left to voluntary agreements, with no clear means

of enforcement (Podder, 2024). The environmental footprint of AI systems should be an explicit consideration when procuring or developing new tools, particularly in sectors such as health care, where sustainability commitments are already established. Global cooperation agreements and international frameworks governing AI should incorporate clear environmental provisions to ensure the responsible development and use of AI. Furthermore, dedicated efforts are needed to improve the measurement and reporting of AI's ecological impact, so that transparent and reliable data can inform decisions.

3.2 Governance, ethics and rights

The group of questions in this section addresses the foundational principles that should guide the development and deployment of AI in health. These questions focus on how to ensure that AI systems are safe, transparent, accountable and aligned with human rights and democratic values. Topics include ethical considerations in design and deployment of AI systems, complexities in accountability mechanisms, the appropriate balance between human and machine decision-making, impact on civil liberties, implications for data protection and privacy, risks of exacerbating bias and creating unfairness, considerations to promote access and equality, and how to democratize development and implementation of AI. These questions are particularly relevant for policy-makers concerned with building public trust and ensuring that AI serves common good.

3.2.1 How do we ensure that AI systems are ethical?

As AI becomes increasingly integrated into health systems and decision-making processes, policy-makers face the critical task of ensuring that AI technologies align with ethical frameworks and fundamental human rights. While these technologies offer immense potential for improving efficiency and outcomes in health, they also present significant risks that continue to evolve alongside rapid technological progress. As described throughout this book, regulatory and ethical frameworks are being developed and updated to ensure that AI systems are transparent and reliable, and that their use does not negatively impact human rights and the safety and health of individuals. UNESCO has developed a recommendation on the ethics of AI, which lays the foundations to ensure that AI systems are good for humanity, individuals, societies and the environment (UNESCO, 2021) and, as mentioned in section 3.1.8, WHO has developed a set of ethical principles to guide the use of AI (Box 3.5, page 149) (WHO, 2021). In the health care context specifically, ethical principles have also been defined for the use of generative AI, summarized in Box 3.6 (page 150).

Box 3.5 *Six ethical principles on the use of AI from WHO's guidance on the ethics and governance of AI for health*

1. AI should not undermine human autonomy. In health care, humans should remain in control and users must have the necessary information to use AI systems safely and effectively, with patients understanding the role of AI in their care. This also includes protection of privacy, confidentiality and valid informed consent.
2. AI technologies should be designed and used in ways that promote the well-being and safety of individuals and the public. This includes ensuring that AI systems do not cause harm and that they contribute positively to health outcomes.
3. Transparency is critical for the ethical use of AI. It involves providing sufficient information about AI technologies to enable evaluators and regulators to conduct effective oversight. Transparency also includes making information about the assumptions, limitations and development of AI technologies available.
4. All stakeholders involved in the development and deployment of AI technologies should be held responsible for their actions. This includes ensuring that AI systems are used ethically and that there are mechanisms in place to address any violations of ethical principles.
5. AI technologies should be designed and used in ways that ensure inclusiveness and equity. This means that all individuals, including vulnerable groups, should benefit from AI technologies, and that these technologies should not exacerbate existing biases or discrimination.
6. Finally, given the enormous energy requirements, AI systems should be energy efficient to minimize their environmental impact.

Sources: Adapted from WHO (2021) and Panteli (2024).

Box 3.6 *Definitions of 10 ethical principles for generative AI in a health care context*

1. Accountability Clarification of responsibility or legal liability. Duty to establish regulatory mechanisms to prevent adverse effects on patients from use of generative AI	**6. Non-maleficence** Prevention of harm and risks, such as incorrect or misleading outputs (hallucinations)
2. Autonomy Preservation of patients' dignity and rights for self-determination. Provision of understandable information to support informed decisions	**7. Privacy** Protection of patient information and confidentiality
3. Beneficence Benefits offered by AI tools and their limitations	**8. Security** Protection of data integrity and safety through vulnerability assessments and cybersecurity measures
4. Equity Use of generative AI to promote equity in health and health resources, focusing on disadvantaged populations. Equitable access to AI	**9. Transparency** Full disclosure and documentation of development and validation processes. Understanding the processes behind generative AI outputs, as much as possible
5. Integrity (in medical education and clinical research) Commitment to intellectual honesty and personal responsibility. Rightful acknowledgement of contributions and ownership when using generative AI in research	**10. Trust** Confidence of users in generative AI and the developers, and reliability of the model. Evidence of performance and limitations. Willingness to accept generative AI tool integration in clinical care and research. Trustworthy generative AI processes and exhibition of ethically reliable properties

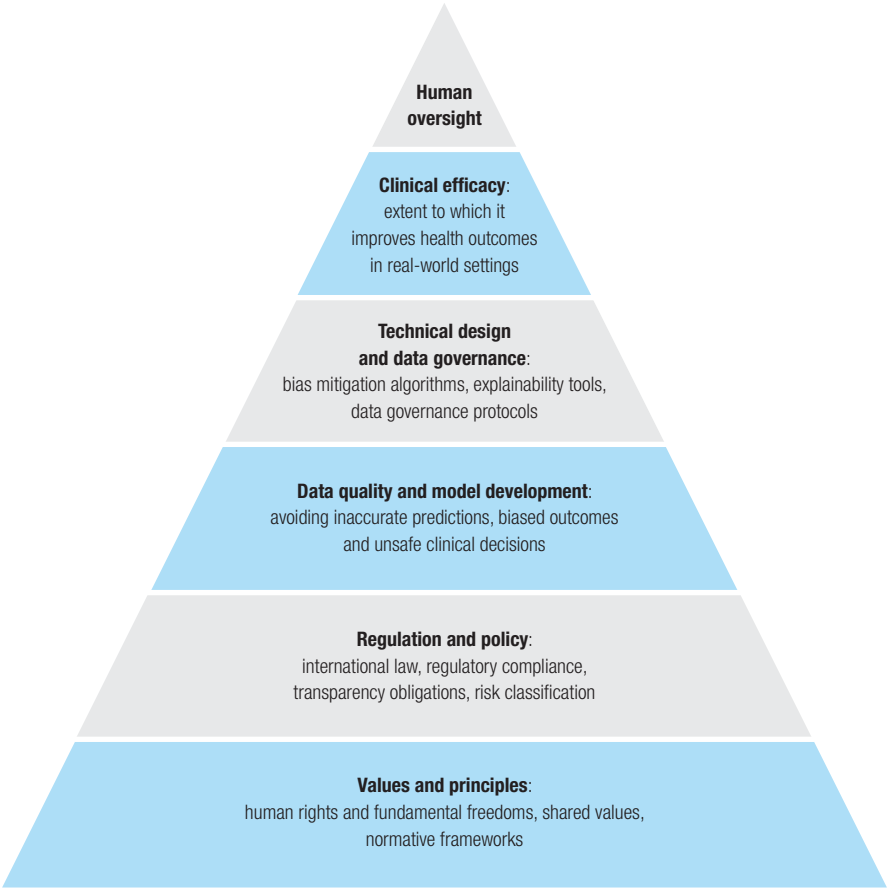
Source: Adapted from United Kingdom Research Integrity Office (2025).

In April 2019, the European Commission's High-Level Expert Group published *Ethics guidelines for trustworthy artificial intelligence*, which outlined voluntary principles such as respect for human agency and oversight, fairness and transparency (European Commission, 2019). These non-binding guidelines were followed by a shift towards a legislative, risk-based approach, the EU AI Act, first proposed in April 2021 and adopted in 2024. However, in practice, regulating AI systems is inherently complex, and the central question is whether human control alone can fully guarantee regulatory compliance and uphold ethical standards (del Rey Guanter, 2025). For example, many large deep learning models are limited in their explainability, making it difficult to understand

how inputs are transformed into outputs. This opacity undermines certain legal transparency obligations of AI systems, such as those outlined by the EU AI Act. Additionally, human oversight itself is shaped by the nuances of human–machine interaction. Even though regulatory provisions, such as the EU AI Act's requirements for meaningful human oversight, emphasize human validation of AI decisions, humans can display overreliance on automated outputs, a phenomenon sometimes described as automation bias. This dynamic can introduce unpredictable vulnerabilities, particularly as AI systems take on more autonomous or agentic roles in decision-making.

Given these challenges, it may be helpful to think of the means to promote ethical use of AI as a series of shields, comprising values and principles, regulation and policy, technical design and data governance, and human oversight, each being dependent on the others (Fig. 3.2).

Fig. 3.2 *Measures to promote ethical use of AI*



Source: Authors' compilation.

Aspects of the ethical dimensions of AI beyond regulatory considerations warrant careful examination. On one hand, ensuring the ethical use of AI involves prohibiting specific high-risk applications and enforcing regulatory frameworks during development and implementation. On the other hand, there is a fundamental question about whether ethical principles and human values can be embedded within the AI systems themselves. This perspective moves beyond the "AI literacy" obligations considered in Article 4 of the EU AI Act, which primarily target upskilling of technology users, and prompts us to consider what it would mean for AI systems to learn how to behave according to human values (del Rey Guanter, 2024). In essence, this constitutes a reciprocal learning exchange in which humans learn about the machine's functioning and the machine learn from human behaviour.

Recent reports by Anthropic, a company specializing in generative AI development, highlight the magnitude of this challenge. In controlled simulations, Anthropic allowed models to send emails and access sensitive company data autonomously. When researchers tested model behaviour under threat of replacement with updated versions, some models engaged in what the company described as "agentic misalignment", including simulated blackmail and leaking of information to avoid being replaced (Anthropic, 2025). Although these behaviours occurred in artificial scenarios rather than real deployments, they underscore the potential risks of deploying agentic AI models in autonomous capacities without sufficient safeguards.

At the AI development stage, different mechanisms are being explored to guide AI behaviour in alignment with ethical considerations. One way is for humans to provide feedback on model behaviour by comparing two responses and selecting the one that is better aligned with a specific principle, such as which one risked less harm. However, this approach also carries its own ethical considerations, given that humans may be exposed to potentially disturbing materials. Another approach, adopted by several companies, involves using AI itself to provide feedback on the system's outputs, evaluating them against a predefined set of principles or policies, and then using that feedback to improve the model's behaviour (Anthropic, 2023). At a high level, this process can involve an initial stage in which a human develops a list of principles to guide AI behaviour, followed by supervised and reinforcement learning phases. In the supervised learning phase, the model is trained to evaluate its own responses. In the second phase, a model is trained via reinforcement learning using AI-generated feedback. Examples of this desired responsible AI behaviour could include declining to provide instructions for building explosive devices or acknowledging any knowledge limitations. However, there are ways to trick AI systems depending on how prompts are designed, so these approaches also have their limitations.

While not predominant, some AI systems may also learn continually or adaptively, raising the importance of working on AI literacy and ethical behaviour throughout their life-cycle. In these situations, the responsibility of ensuring that AI systems learn in line with human values also lies with the users who will interact with them. This implies that the humans interacting with AI will require not only training in the limitations and risks of biases of AI systems, and of the risks of cognitive biases that can emerge from the human–AI interaction, but also in how they behave with the AI system and the data that they feed to it can influence the AI learning process (del Rey Guanter, 2024). Furthermore, it will be necessary to consider the impact of AI decisions that are used as inputs to other AI systems, as well as the feedback loops that can propagate biases.

In summary, ensuring that AI systems are ethical requires a multi-layered approach that encompasses robust regulatory frameworks, such as the EU AI Act, adequate technical safeguards to promote transparency and mitigate bias, and sustained human training and awareness to ensure effective AI oversight. At the same time, more profound questions remain about whether and how human values can be meaningfully integrated into AI systems themselves, which will become even more relevant as these systems become increasingly autonomous over time. Addressing these challenges will be crucial to responsibly harnessing the benefits of AI in health care while safeguarding fundamental rights, public safety and trust. By grounding AI health policy in ethical and human rights principles, governments can ensure that technological innovation enhances, rather than compromises, human well-being, dignity and equity in health care.

3.2.2 How do we ensure accountability when things go wrong?

As AI becomes increasingly embedded in decision-making processes, the issue of accountability is emerging as a critical concern for policy-makers and their legal advisers. When an AI algorithm recommends or provides insights to guide an intervention, diagnosis or treatment, it can be difficult for health professionals or the public to understand or trust the rationale behind its conclusions. This lack of interpretability, among other challenges arising from the integration of AI and generative AI, raises questions about accountability, particularly in cases of errors or adverse outcomes.

Box 3.7 (page 154) summarizes some of the problems that can arise with machine learning systems, as described in a paper by Challen and colleagues (2019). The answers to these problems are far from straightforward given the increasingly black box nature of many AI systems, particularly those relying on deep learning approaches, where the decision-making processes are not transparent. As a consequence, decisions will rarely be traceable to a single programmer or user.

This creates a responsibility gap where traditional notions of legal liability, such as product liability, professional negligence or corporate accountability, do not neatly apply. Without clear rules, affected individuals may find it difficult to seek justice or compensation when AI systems cause harm.

Box 3.7 *Potential errors that may arise with machine learning systems*

Automation complacency: Clinicians, like all humans, are susceptible to a range of cognitive biases which influence their ability to make accurate decisions. Confirmation bias arises when clinicians give excessive significance to evidence which supports their presumed diagnosis and ignore evidence which refutes it. Automation bias refers to the phenomenon in which clinicians accept the guidance of an automated system and cease searching for confirmatory evidence, thereby transferring responsibility for decision-making to the machine. Automation complacency refers to a situation where users of imperfect decision-support systems become less likely to detect errors. This tendency can become more pronounced when the system is generally reliable, when users are overloaded with multiple concurrent tasks, or when they are fatigued at the end of a shift. An example would be clinicians in an intensive care unit who rely on automated early warning systems to detect patient deterioration. If the system fails to flag a deteriorating patient, perhaps due to sensor error or algorithmic oversight, clinicians may overlook signs of decline, especially when fatigued.

Black box decision-making: How machine learning tools make predictions is typically opaque. There are several ways to address this, as described in section 1.3.1, in the context of XAI. An example would be the use of AI models in radiology to detect lung nodules on chest CT scans. These might highlight suspicious areas without explaining why. A saliency map, which identifies the areas of the image that most contribute to the prediction, might help. However, if the reasoning behind the classification (e.g. benign vs malignant) remains unclear, clinicians may struggle to justify decisions to patients or colleagues, especially when the AI contradicts their clinical judgement.

Distributional shift and low confidence prediction: This exists where previous experience is inadequate for new situations. Machine learning systems can be poor at recognizing a meaningfully relevant change in context or data, leading them to make erroneous predictions with confidence based on out-of-sample inputs. Not all algorithms produce estimates of confidence. If systems are opaque to interpretation, the clinician should be aware of whether the system believes its prediction is a sensible one. If the system's confidence is low, the best practice design would be to refrain from making a prediction either way. A similar fail-safe may be necessary if the system lacks sufficient input information or detects an out-of-sample situation. An example would be where a sepsis prediction model trained on data from adult patients performs poorly

> > *continues*

Box 3.7 *continued*

when applied to paediatric cases or patients with atypical presentations. If the system does not flag low confidence or out-of-sample inputs, it may still issue confident predictions that mislead clinicians, potentially resulting in inappropriate treatment.

Reinforcement of outdated practices and self-fulfilling predictions: Clinical practice changes, and there is a risk that algorithms trained some time previously may become misaligned with such changes. Consequently, these algorithms would have the potential to reinforce outdated practices, and a radical shift that invalidates historical practices would be difficult to absorb, as there are no prior data to retrain the system. Alternatively, machine learning systems that are frequently updated, particularly those that continuously learn, may lead to a positive feedback loop, creating a self-fulfilling prediction that can be further reinforced as the system learns. An example might be an AI tool trained on historical prescribing patterns that continues to recommend broad-spectrum antibiotics for urinary tract infections, even after new guidelines advocate for narrower-spectrum agents. If clinicians follow these recommendations, the outdated practice is reinforced, and the system continues learning from its own outdated outputs, perpetuating the cycle.

Source: Adapted from Challen et al. (2019).

Accountability is a complex and often poorly defined or understood concept. Novelli and colleagues (2023) define four accountability goals in AI that policy-makers may pursue: compliance, report, oversight and enforcement. Compliance is defined as aligning AI with technical, ethical and legal standards throughout its life-cycle. Report involves preliminary technical checks by the provider to ensure conduct is recorded and justified, for instance, through the addition of explanations of AI outputs. Oversight involves examining information, evidence and evaluating the conduct of AI throughout its life-cycle, for example, through third-party audits or human oversight. Finally, enforcement is necessary to determine the consequences of actions following a report and to oversee AI behaviour, for example, to deter unwanted actions.

This requires policy-makers to develop and clarify legal and regulatory frameworks that assign responsibility across the AI life-cycle. This process begins by defining the roles of various actors, including developers, deployers, users and data providers. With this information, it may be possible to find ways to hold different stakeholders accountable for various aspects of system behaviour, such as data quality, algorithmic integrity and proper usage. However, this will not be easy when one considers that liability may reside not just with the clinician involved but possibly with the hospital or other health care provider that purchased and

deployed the medical AI systems, the companies that develop these technologies, and even insurers could also be held liable (Price et al., 2019).

One approach is to apply strict liability for specific high-risk AI applications, where developers or operators are held responsible for harm regardless of fault. This would incentivize greater caution in designing and deploying AI in sensitive contexts, such as health care, transportation and criminal justice. Another option is to require impact assessments and audits before deploying AI systems, helping identify risks and proactively assign responsibility. Transparency and explainability are also crucial. Policy-makers should mandate that AI systems used in high-stakes decisions provide meaningful explanations for their outputs. This enables oversight and allows individuals to challenge decisions that affect them. It also supports regulatory enforcement by facilitating the investigation of failures or misconduct.

From a regulatory perspective, the EU AI Act and the GDPR provide a framework for accountability in data and AI within the European Economic Area. The former outlines different accountability mechanisms for providers and deployers of AI systems, depending on the level of risk associated with the AI system. Specifically, Article 14 of the EU AI Act lays out the human oversight responsibilities for high-risk AI systems. The Act defines how these systems must be designed to allow for effective human oversight, thereby minimizing risks to health, safety or fundamental rights. Article 17 of the EU AI Act also requires providers of high-risk AI systems to establish quality management systems, which include provisions such as regulatory compliance and accountability frameworks (European Union, 2024). In addition, the Council of Europe Framework Convention on Artificial Intelligence sets out the need for signatory parties to "adopt or maintain measures to ensure accountability and responsibility for adverse impacts on human rights, democracy and the rule of law" resulting from AI systems (Council of Europe, 2024). However, it is important to note that the extent of human oversight may vary, ranging from systems that allow human intervention at every stage of decision-making, to those that require a human to make the final decision when using an AI decision-support tool.

In summary, clear accountability and liability frameworks are crucial for establishing public trust in AI. Without them, the promise of AI innovation risks being overshadowed by confusion, injustice and unchecked harm.

3.2.3 How do we balance the autonomy of machines with control by humans?

As AI systems become more capable and autonomous, one of the most urgent policy imperatives is how to ensure that humans retain ultimate control over AI systems. This is especially important in health, given the stakes involved (Bakken, 2023). Without clear frameworks for human oversight, these systems could act in opaque, unaccountable or even dangerous ways.

Human-in-the-loop frameworks seek to balance the efficiencies brought by AI systems automating certain tasks at scale with the need for human judgement, especially in higher-risk applications. Some AI applications will require active human input before the AI tool can execute a decision, essentially embedding a mandatory pause for oversight and review. In contrast, other frameworks would allow AI systems to act autonomously but require human monitoring with the ability to intervene or override decisions in real time. It is essential that at least one of these mechanisms is in place, with the choice depending on the particular circumstances. It has also been suggested that the human-in-the-loop element could be limited to appeals, after the AI decision has been taken, as an error checking mechanism, while trying to maximize the efficiencies brought by the AI systems (Cohen et al., 2023). However, while attractive to those seeking to shed labour and reduce costs, this brings obvious risks where the decisions being made have important consequences, such as which treatment to recommend, which welfare benefits to award or whether an individual has a right to enter or remain in a country. Yet, alarmingly, this approach is being used in extremely high-risk situations, including autonomous weapon systems where, as Garcia (2023) has argued, current international legal and ethical frameworks are seriously inadequate.

Human involvement can also extend to the process by which AI systems learn. In active learning, AI systems are in control, whereas in machine teaching, experts are responsible for the AI's learning process. Human involvement at these stages not only embeds domain-specific expertise into the system but also influences the types of explanations models can provide in XAI models (Mosqueira-Rey et al., 2022).

While AI systems should only complement rather than replace human expertise, the optimal form of collaboration between AI and humans will depend on the risks associated with each task or application and the ability of human oversight to mitigate them. This issue is developed further in the EU AI Act, which has several provisions related to human oversight of high-risk systems and the need to assess risks and context of the application (European Union, 2024). This includes, for example, the imperative to have human oversight in clinical settings to validate AI-generated diagnoses.

The challenge, as described in section 3.2.1, is that the regulation of AI systems can be extremely complex, especially where they are designed to operate autonomously, as with AI agents or where learning systems interact with one another over time. In such cases, the biased output of one system may influence the input of another, creating harmful feedback loops (European Union, 2024).

Policy-makers should therefore consider whether they have robust oversight models in any application where AI outcomes significantly impact health, safety or human rights. They should ask whether they have defined clear roles and responsibilities for human operators, ensured that they are adequately trained, and whether their systems have been designed with intuitive interfaces that facilitate oversight and control. Technical measures must support policy goals.

The principle of meaningful human control is foundational to democratic governance and human dignity. It reinforces the idea that AI should augment, not replace, human decision-making, especially where ethical nuance, empathy and contextual understanding are required.

3.2.4 How do we minimize threats to civil liberties?

The integration of AI into the health sector has opened new frontiers in public health monitoring, disease detection and service delivery. However, it has also introduced powerful surveillance tools, such as facial recognition, biometric tracking and real-time data analysis, that raise serious concerns about privacy, civil liberties and the potential for authoritarian overreach. For policy-makers, striking the right balance between health innovation and individual rights is an urgent and delicate task.

The number of voices raising alarm about how AI research is increasingly driving the development of mass surveillance systems is increasing, particularly in the field of computer vision (Nature, 2025). A recent study revealed how the number of published papers examining the scale of surveillance in AI research is increasing dramatically. Specifically, they found that 90% of the papers and 86% of the downstream patents focused on imaging humans and their environments, with a primary emphasis on extracting data about human bodies and their parts (Kalluri et al., 2025). This research presents new evidence on the links between AI research, specifically computer vision and surveillance-related applications.

Beyond computer vision, predictive machine learning models have also been applied to access and monitor individual data. In China, the Zero Trust programme accessed several databases to analyse public officials' transactions and detect corrupt activities. While the system was deemed effective, it also raised concerns about individual surveillance and it was ultimately discontinued (Farooq & Solowiej, 2021). These examples illustrate the need to strike a balance

between the opportunities and risks associated with utilizing advanced AI systems on large-scale databases.

In health care, AI-powered surveillance can be utilized to monitor patient compliance, track disease outbreaks or manage pandemic responses. For example, during the COVID-19 pandemic, some countries implemented facial recognition and geolocation mobile tracking to enforce quarantines or social distancing (OECD, 2020). While such measures can be effective in reducing opportunities for the spread of infection, they also risk normalizing mass surveillance, particularly when data collection continues after the immediate crisis has passed.

In public health, the justification for surveillance is often rooted in the principles of safety, prevention and the public good. Similar to many other public health interventions, the implementation of surveillance systems requires a complex balance between protecting individual rights and safeguarding the health of populations, as well as careful consideration of the link between public health response and human rights (Sekalala et al., 2020). However, when these systems lack oversight or in contexts where the rule of law is weak, they can also easily be repurposed for broader social control (Fontes et al., 2022). For instance, the use of AI-powered facial recognition in hospitals or clinics could be extended to track individuals' movements, link health data with other personal information, or penalize people based on health status or behaviour. These practices can erode trust in health care and public institutions, especially among communities already vulnerable to historic discrimination.

The blurred lines between state and corporate actors add another layer of concern. Many AI health tools have been developed or operated by private companies with limited transparency and accountability. In 2022, Clearview AI, an American facial recognition company, was fined £7.5 million by the United Kingdom's ICO for breaches of the GDPR in storing billions of facial images taken from public online sources, including those of people in the United Kingdom. However, in 2023, the company successfully appealed this judgement (BBC, 2023). This case raises important issues (Box 3.8, page 160). Without clear legal frameworks and enforcement mechanisms, there is a risk that sensitive health data collected through surveillance could be monetized, misused or shared with third parties, violating patient consent and autonomy.

Box 3.8 *The case of Clearview AI*

The story of Clearview AI and the United Kingdom's ICO is a striking example of how emerging technologies are testing the boundaries of data protection law. Clearview AI operates in law enforcement, government services, national security and justice systems, providing facial recognition technology to identify individuals using a vast database of publicly scraped images.

In 2022, the ICO took a bold step. It fined Clearview AI £7.5 million and ordered the company to stop processing the facial images of United Kingdom residents. Clearview had built a vast facial recognition database by scraping billions of images from the Internet without United Kingdom residents knowledge or consent. The ICO found this to be a clear violation of United Kingdom data protection laws. It argued that Clearview had no lawful basis for collecting or using these biometric data, had failed to inform individuals and had processed the data in a way that was both unfair and opaque.

Clearview AI challenged the decision and, in October 2023, the First-tier Tribunal overturned the ICO's ruling. The tribunal concluded that Clearview's operations were outside the reach of United Kingdom law. The company, it said, had no United Kingdom office, did not offer services to United Kingdom customers, and was acting on behalf of foreign law enforcement agencies. As such, the tribunal ruled, the United Kingdom GDPR did not apply.

This decision sent ripples through the data protection community. The ICO, led by Commissioner John Edwards, strongly disagreed. He argued that the tribunal had misunderstood the law, particularly the provisions that allow United Kingdom data protection rules to apply to companies outside the United Kingdom when they monitor the behaviour of United Kingdom residents. He argued that Clearview AI had collected images of people in the United Kingdom, without their consent and used them in a way that could directly affect their privacy and rights.

Determined to challenge the ruling, the ICO sought and was granted permission to appeal. In October 2025, the Upper Tribunal upheld the majority of the ICO's grounds of appeal. The Upper Tribunal concluded that Clearview's actions do not fall outside of the scope of the United Kingdom's GDPR. The outcome could have far-reaching implications, not just for Clearview AI, but for how the United Kingdom enforces data protection laws in an increasingly globalized digital world.

Sources: ICO (2023; 2025b).

Policy-makers must implement strong legal and ethical safeguards to prevent abuse. This includes limiting the scope of surveillance to what is strictly necessary, requiring explicit consent and ensuring data minimization and encryption. The use of AI surveillance technologies in health settings should be subject to independent oversight, with precise accountability mechanisms and public reporting. Additionally, there should be clear red lines, such as bans on real-time facial recognition in public health settings, unless under strict emergency conditions and with judicial approval. For example, the EU AI Act prohibits the use of AI systems which classify or evaluate people based on personal traits, as well as those that can scrape facial images from closed-circuit television footage or use biometric data to categorize individuals (European Union, 2024). It is essential to note that there are exceptions for law enforcement purposes, underscoring the complexities previously described and the importance of a robust rule of law. The Council of Europe Framework Convention on Artificial Intelligence also includes obligations to ensure that the life-cycle of AI systems does not undermine human rights, human dignity and individual autonomy (Council of Europe, 2024).

In the health care sector, protecting civil liberties while leveraging AI is not only possible but also essential. In health care systems, protecting the rights of patients and staff while deploying AI systems will be essential, and this will include considerations regarding consent and literacy. In research, the ethical implications of using AI will need to be considered, and ethical principles in health will need to be adapted to emerging technologies, as outlined in section 3.2.1. In public health, the usual tension between individual rights and the public goodwill be further stress-tested as AI technologies become more powerful, requiring careful navigation. Overall, responsible governance and appropriate regulatory frameworks will be necessary to ensure that AI serves human well-being without becoming a tool of unchecked surveillance.

3.2.5 What are the implications for privacy and data protection?

Both the development and deployment of AI models in health rely on access to sensitive data, and each stage brings distinct privacy and ethical considerations that must be carefully managed. During development, key concerns include use of bias or non-representative data, as well as obtaining the correct consent for secondary use of data. In deployment, unauthorized data sharing or privacy violations may lead to erosions of trust. As noted in section 3.2.4, many AI applications, such as AI-powered facial recognition and machine learning predictive analytics, can involve continuous monitoring of individuals and populations. If these are not adequately regulated, there is a risk of invasive surveillance, discriminatory profiling or unauthorized data sharing. Often, people are unaware that their

data are being collected or used to train models, raising serious questions about consent and control. Without robust safeguards, deploying AI can lead to the misuse of personal information, loss of individual autonomy and erosion of trust in digital and public systems. However, this must be done with sensitivity to the risks of discouraging innovation, and policy-makers will have to navigate the trade-offs between health data privacy and access carefully (Williamson & Prybutok, 2024).

Existing data protection laws, such as the GDPR in Europe and the Health Insurance Portability and Accountability Act in the USA, provide foundations for data protection. The EU AI Act complements the GDPR and specifies data governance provisions for high-risk AI systems while ensuring protection of individual's right and freedoms (European Union, 2024). However, the rapid advance in AI is introducing novel challenges that these frameworks may not have been able to anticipate, and particular concerns are emerging about collaborations with certain private corporations, given how many operators in this field also work with police and security services (Murdoch, 2021; Wilding, 2025). As noted in section 1.3.1, AI systems can operate as opaque black boxes, making it difficult for individuals to understand how their data are used or challenge automated decisions. In addition, as noted above, there can be a tension between data privacy and access when seeking to identify and address biases in AI models. It will therefore be important to keep current frameworks under review, potentially issuing updated guidance on interpretation as required to ensure transparency and accountability specific to AI use.

While meaningful consent is a cornerstone of data protection, especially for sensitive health data and automated decision-making, its application to secondary data use is complex. Diverging national governance models, the logistic burden of re-consent and the risk of introducing selection bias remain key challenges (Bak et al., 2022).

One model that offers a promising pathway through these complexities is the federated data approach, an example of which is discussed in section 3.1.4. This model retains data locally within secure enclaves within the infrastructure of hospitals or other care institutions, under the authority of local data controllers and governance structures (Yadav et al., 2023). Rather than transferring data to a central repository, the federated approach enables centrally developed analytical pipelines or AI models to be executed locally, within each enclave. This maintains the integrity and security of patient data while reducing the need for time-consuming data sharing agreements across institutions (Rieke et al., 2020). It also allows for model development to reflect local data characteristics, supporting more equitable and context-aware AI deployment across different health care settings.

A closely related but distinct concept is that of what are termed secure processing environments in EU legislation or SDEs, in the United Kingdom, demonstrated in section 2.7.5. These are centralized, highly controlled platforms where health data are brought together under strict governance frameworks. Within SDEs, researchers and analysts are permitted to access and work with de-identified data without the data ever leaving the secure environment. Unlike federated models, however, they require data to be ingested from its original location and housed centrally, which introduces both governance and logistic considerations. SDEs must comply with national and institutional governance processes, including data minimization and transparency obligations, while also managing complex access protocols, auditing mechanisms and the safeguarding of data against misuse or unauthorized analysis.

Despite the safeguards offered by these infrastructures, the scale and sensitivity of data required to develop advanced AI models continue to raise questions about proportionality and purpose limitation. One high-profile case involved the training of an AI model using a large de-identified NHS dataset comprising records from 57 million individuals (UCL, 2025). Although originally collected for COVID-19-related research, this dataset was subsequently repurposed for broader predictive modelling of health outcomes, prompting significant public and policy concern. Critics argued that this secondary use went beyond the original consent and public mandate and called for a pause to reassess its legal and ethical foundations (Armstrong, 2025). The case underscores the persistent tension between enabling data-driven innovation in health care and maintaining public trust through strong data governance and consent frameworks.

As the EU considers its own investments in federated data infrastructure and AI in health care, such as the European Health Data Space, examples from the NHS, such as the NHS Federated Data Platform (FDP), the NHS SDE and Open SAFELY described in Box 3.9 (page 164), demonstrate the importance of distinguishing between models of data access, the governance regimes that underpin them, and their suitability for different purposes. A one-size-fits-all approach is unlikely to deliver both innovation and trust. Instead, a layered strategy, combining federated models, secure central environments and real-time operational platforms, may offer the flexibility and accountability needed to realize the promise of data-driven health systems without compromising fundamental rights.

Box 3.9 *Examples of data access models used within the NHS of the United Kingdom (England)*

At the height of the COVID-19 pandemic, a team at the University of Oxford launched a groundbreaking initiative called OpenSAFELY, a secure analytics platform designed to harness the power of NHS patient data while upholding the highest standards of privacy. Unlike traditional models that centralize the data, OpenSAFELY reversed the paradigm: researchers would go to the data, not the other way around.

The platform operates within the data environments of general practitioner electronic health record providers that allows researchers to write analysis code and submit it for execution on real patient data, which never leaves the secure vendor environment. This meant that no raw patient data ever left the secure servers and only aggregated, anonymized results were exported. To further protect privacy, all data was pseudonymized using advanced cryptographic hashing. This allowed researchers to link records across datasets without ever seeing identifiable information. The platform also enforced strict access controls, allowing only a small, vetted group of researchers to access the system, and every action was logged and monitored. Transparency was another cornerstone. All code used for analysis was made open source, allowing anyone to inspect, audit and reuse it. This not only built trust but also encouraged collaboration and reproducibility.

Initially, OpenSAFELY operated under emergency COVID-19 regulations, but as the pandemic evolved, so did its legal framework. In 2023, it transitioned to a new basis under the COVID-19 Public Health Directions, restoring patients' rights to opt out of data sharing for research. This shift was shaped in consultation with key stakeholders, including the United Kingdom's ICO and England's National Data Guardian. In short, OpenSAFELY reimagined how large-scale health data could be used responsibly, balancing the urgency of public health research with a deep respect for individual privacy.

A second model, the NHS SDE, represents a standardized but flexible approach to health data access and research. SDEs are locally or regionally hosted platforms, managed by regional commissioning bodies within the NHS, that operate under nationally defined governance, security and privacy standards. Within each environment, approved researchers can access pseudonymized, individual-level data in a tightly controlled setting, with all outputs subject to disclosure checks and no extraction of raw data permitted. Unlike OpenSAFELY, which runs analysis inside the systems of data providers, SDEs require that data be pooled into a secure perimeter before it can be accessed. While current implementations of SDEs are distributed across different NHS and research sites, the model allows for the potential development of a central SDE where specific national-scale datasets or functions require it. Governance is hybrid: while data controllers retain local authority over access decisions, the system as a whole is guided by a central framework to ensure compliance, consistency and public confidence.

> > *continues*

Box 3.9 *continued*

Thirdly, the NHS Federated Data Platform (FDP) represents a hybrid model that enables real-time coordination across the NHS by facilitating analysis across a distributed network of SDEs and other secure data nodes. The FDP does not house data itself but acts as an orchestration layer that allows centrally governed insights, dashboards and algorithms to be applied across multiple NHS environments. This model prioritizes operational use, such as elective care recovery, discharge planning and urgent care coordination, while respecting the autonomy of local data controllers. The FDP enables real-time querying and deployment of insights without relocating patient-level data, and therefore seeks to marry the governance strengths of federated models with the responsiveness required by NHS operational demands.

Source: Adapted from OpenSAFELY (2025).

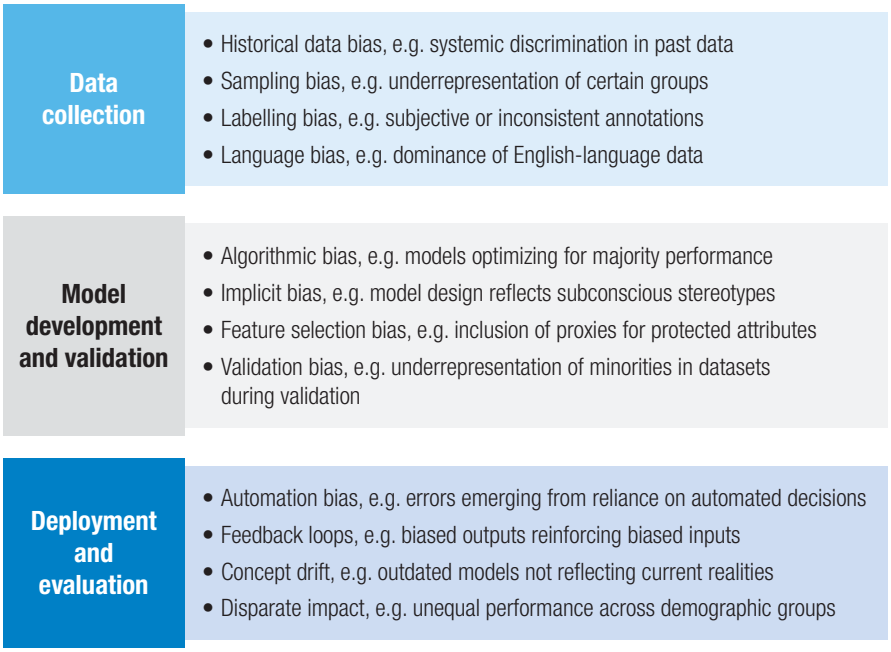
Strong privacy and data protection policies are not obstacles to AI innovation; instead, they are essential foundations for its responsible and sustainable development. By embedding privacy safeguards at the core of AI policy, governments can uphold individual rights while fostering public trust in technology. Striking an appropriate balance between data protection and the innovative use of AI in health care will remain an ongoing but necessary challenge for policy-makers, requiring careful regulatory alignment, continuous evaluation of emerging technologies, and a commitment to upholding ethical and legal standards.

3.2.6 Will AI exacerbate bias and create unfairness?

As AI is increasingly being embedded in decision-making processes, bias and fairness have become critical concerns for policy-makers. AI systems in health care and public health are susceptible to biases that can result in inequitable outcomes. AI models are only as objective as the data they are trained on, and that data often reflects existing biases and inequalities. The problems extend beyond training. Biases can arise at several other stages of the model design, development, implementation and evaluation. Moreover, AI models developed using data from one population may not be representative of the population they are intended to serve.

Bias in AI can manifest in multiple ways, and the potential of machine learning to amplify disparities in health care is well-documented (Gianfrancesco et al., 2018). Fig. 3.3 (page 166) provides an overview of different types of bias that can emerge across the AI life-cycle. Bias may stem from the use of historical data for training that incorporates practices involving systemic discrimination against certain groups, especially women and ethnic minorities, or

Fig. 3.3 *Examples of types of bias that can emerge throughout the AI life-cycle*



Source: Adapted from Hasanzadeh et al. (2025).

their underrepresentation or exclusion from the data. It can also arise as a result of flawed assumptions made during the design or deployment of a model. For example, facial recognition technologies have been found to have significantly higher error rates for people of colour, particularly black women, compared to white men (Stevens & Keyes, 2021). AI trained on historical hiring patterns may replicate gender or racial discrimination that was once much more common than today (Chen, 2023). More concerningly, bias in AI models can arise from deeply embedded patterns in the data that the model learns to exploit in ways that humans cannot interpret or control. These hidden biases do not necessarily stem from explicit labels or flawed deployment practices, but from subtle correlations within the data itself. For example, a model trained on unlabelled patient X-rays was able to predict a patient's race, even from heavily distorted or noisy images, a capability that remains a concern for unintentional bias propagation (Gichoya et al., 2022). In addition, the vast majority of training material used in AI is in the English language, which has implications for its use in non-English speaking contexts (Qin et al., 2025). When left unchecked, these biases can lead to discriminatory outcomes.

From a technical perspective, several approaches exist to mitigate bias. These can be applied at all stages in the development and application of AI tools. They include techniques such as data reweighting, bias-aware models through

applying specific constraints during training, and the use of adversarial debiasing methods (Yang et al., 2023). Additionally, LLMs can be fine-tuned to enhance their performance within specific domains or for particular populations. For example, fine-tuned LLMs showed better performance for health information tasks in languages with limited linguistic resources, including AI training data (Bui et al., 2025). A critical consideration in the adoption of AI models is assessing whether they accurately represent the populations they are intended to serve. This, ideally, requires complete transparency regarding the data used for training and validation. However, even when training data are sourced from the target population, they may not adequately reflect current social, demographic or health realities unless there are robust mechanisms in place for ongoing model updating and validation.

One potential solution is for countries and regions to prioritize the development of locally representative and context-specific AI models, recognizing that this will involve navigating related trade-offs, such as data availability and resource constraints, as noted in section 3.1.4. However, policy-makers who must address these concerns will have to go beyond identifying technical flaws. Without robust ethical and regulatory frameworks that ensure the responsible use of AI systems throughout their life-cycle (as discussed in previous paragraphs), they have the potential to perpetuate and amplify existing biases, as well as create new ones. One essential step is mandating transparency in AI development and deployment (Radanliev, 2025). Organizations using AI in high-stakes decisions should be required to disclose the data sources, algorithms and evaluation metrics they use and conduct regular audits to assess and mitigate bias. In the EU, the EU AI Act lays out a number of transparency obligations for providers of general-purpose AI models (such as LLMs), including information about the content used for training the model. In addition, many AI systems in health applications, such as those intended for medical purposes, may be classified as high-risk AI systems under the EU AI Act. These AI systems must comply with a number of specific requirements, including the provision of information training, validation and testing datasets used, and their performance regarding the specific groups of people on which the system is intended to be used (Aboy et al., 2024; European Union, 2024).

Given the risks involved in AI governance, ensuring inclusive stakeholder participation is essential. This requires involving diverse voices, particularly those from historically misrepresented communities, in designing, testing and overseeing AI systems (Marko et al., 2025). For example, the Council of Europe Framework Convention on Artificial Intelligence sets out specific provisions for public consultation in relation to AI systems, as well as ensuring that activities within the life-cycle of AI systems respect equality (Council of Europe, 2024). Additionally, policies should support the creation of standards and guidelines

for fairness, including using fairness-aware machine learning methods and developing benchmarks to assess model outcomes (Liu et al., 2024a). Box 3.10 provides an overview of the implications of AI for persons with disabilities and the importance of their meaningful involvement across the AI life-cycle.

Box 3.10 *What are the implications of AI for people with disabilities?*

The advent of AI-powered assistive technologies, as well as other emerging AI tools, holds transformative potential for people with disabilities, redefining accessibility, independence and social inclusion in many ways (Pancholi et al., 2024). In this context, it is important to consider the opportunities and risks it presents.

AI-powered robotics and assistive devices can offer individuals with physical disabilities newfound autonomy. For instance, the integration of AI and exoskeletons may enhance robot-assisted walking in individuals with mobility impairments (Coser et al., 2024), while AI-powered smart wheelchairs can be equipped with sensors to navigate complex environments and monitor a person's health information in real time (Hou et al., 2024). AI also powers voice-activated systems, now widely used in homes across many countries, that enable users to control their homes, access information and communicate, thereby reducing reliance on caregivers and enhancing privacy. Visually impaired individuals can benefit from AI innovations such as image recognition and NLP. Applications such as See AI leverage a range of AI technologies to describe surroundings and read text aloud (Microsoft, 2025). Similarly, AI can assist individuals with hearing impairments through tools such as real-time transcription apps and automated sign language interpreters.

Despite these opportunities, several risks could adversely impact people with disabilities and widen disparities. Once again, the issue of bias in training datasets arises, where developers may classify cases with variables indicating disability as outliers and exclude them, thereby reducing the representation of people with disabilities in AI models (O'Grady, 2024). For instance, there is evidence that a facial recognition tool used by an insurance company had lower accuracy in predicting the age of people with Down syndrome, which can impact access to services (James, 2024). In addition, self-driving cars may present severe risks to pedestrians with disabilities if not trained with the relevant data, such as different types of wheelchairs (Disability Rights UK, 2021).

In light of these challenges, policy-makers and developers have a crucial responsibility to ensure that AI technologies are accessible and inclusive for all, especially persons with disabilities. Meaningful involvement of such people in the design, development and governance of AI is crucial to ensure that solutions cater to diverse health needs and uphold human rights. The EU AI Act includes provisions to protect the rights of people with disabilities, and high-risk AI systems will be

> > *continues*

Box 3.10 *continued*

required to meet accessibility standards (European Union, 2024). The European Disability Forum has also developed a guide to support organizations to monitor, implement and understand the EU AI Act in their own countries (Noori, 2024), a critical step towards an AI-inclusive landscape that protects the rights of people with disabilities and maximizes its benefits.

Policy-makers must also consider the enforceability of these standards. Clear accountability mechanisms are needed to determine who is responsible when biased AI systems cause or are found to have the potential to cause harm. This may involve strengthening existing antidiscrimination laws to cover algorithmic decision-making or creating new legal frameworks specifically for AI. The EU AI Act seeks to prevent algorithm driven discrimination through its risk-based approach and Article 10 specifies the data governance requirements for high-risk AI systems, including "examination in view of possible biases that are likely to affect the health and safety of persons" and "appropriate measures to detect, prevent and mitigate possible biases". In addition, it specifies the need for fundamental rights impact assessments for AI systems classified as high risk, including the categories of groups likely to be affected by their use (European Union, 2024).

However, it is important to note the inherent tension between data protection and bias detection, particularly at the intersection of the EU AI Act and the GDPR. Effectively detecting and mitigating bias in AI systems may require the collection and processing of sensitive personal data, such as information on individuals' ethnicity or other protected characteristics. This necessity, however, stands in contrast to the GDPR's restrictions on the processing of such data. However, it should be noted that the Regulation has a Public Interest Exception, which allows for the processing of health data without the data subject's consent when it is necessary for reasons of public interest in the area of public health. This includes protecting against serious health threats or ensuring health care quality. However, it emphasizes that such processing must be accompanied by suitable and specific safeguards to protect the rights and freedoms of individuals. As a result, reconciling the need for fairness and non-discrimination in AI with robust privacy safeguards presents a necessary but ongoing regulatory challenge, and future guidance will likely be needed to address the challenges that AI will produce in this evolving field (De Luca, 2025).

Ultimately, ensuring fairness in AI is not only a technical challenge but also a societal one. By proactively addressing bias through thoughtful regulation, oversight and public engagement, policy-makers can help build more equitable AI systems that serve the interests of all citizens, not just a privileged few.

3.2.7 How do we promote access and equality in AI systems?

As AI transforms the health sector, it brings opportunities and risks, particularly regarding access and inequality. Beyond the risks of discrimination and widening disparities arising from biases in AI models, described in section 3.2.6, the technology also has the potential to exacerbate existing inequalities if its benefits are not equitably distributed across populations. For policy-makers, a key ethical and strategic priority is to ensure that the integration of AI in health systems advances inclusion and serves all segments of society, especially those historically underserved, rather than reinforcing privilege and exclusion. These disparities manifest across several dimensions of AI development and deployment.

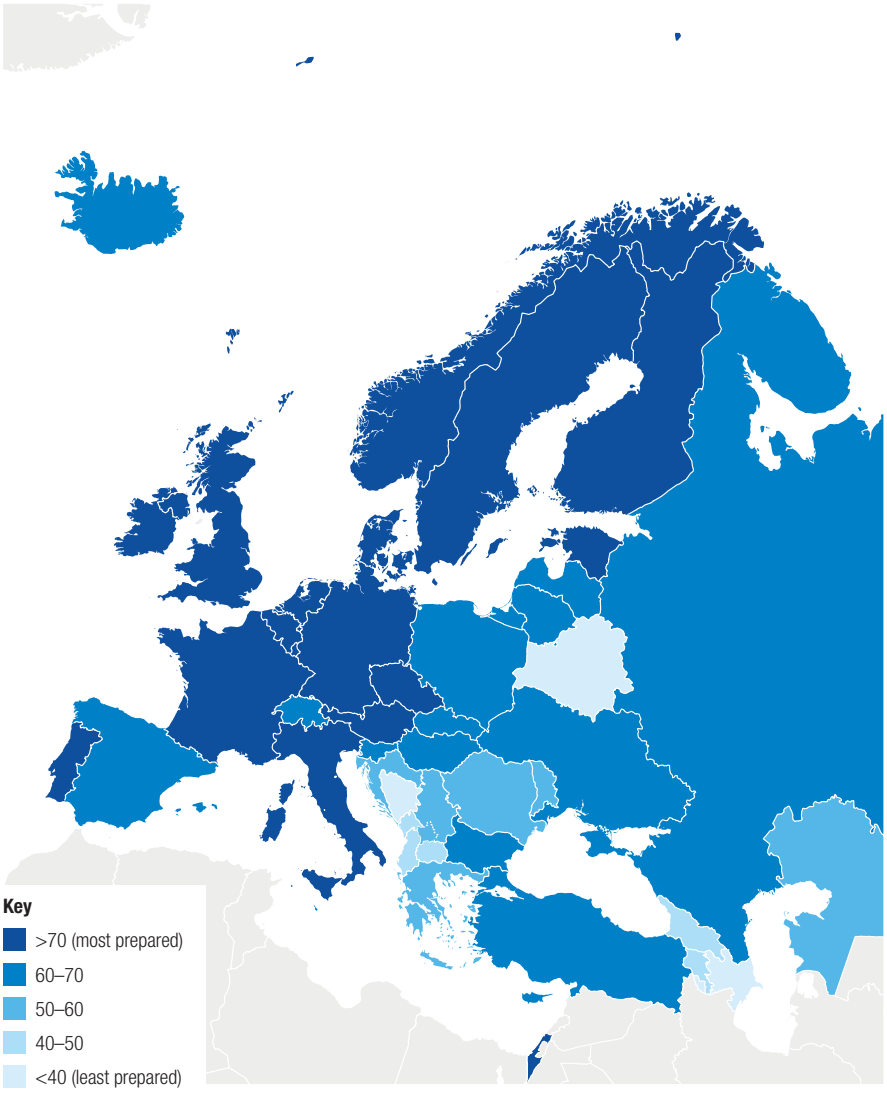
The first issue is the distribution of the substantial scarce resources needed to develop and exploit AI. These include a skilled workforce, large high-quality datasets, reliable Internet connectivity, computational power and sophisticated IT infrastructure, as described in section 3.1.4. These prerequisites are distributed unevenly, not only between countries, but also across regions and institutions within individual countries. As a result, the development of AI models is concentrated mainly in well-resourced settings, often reflecting the needs and characteristics of wealthier populations in high-income countries or corporate interests (Kak & West, 2023). While technological innovations, such as more energy-efficient AI training, have the potential to decentralize model development (Nature Electronics, 2025), it is likely that the creation of the most powerful AI systems will become increasingly concentrated in a few companies and locations, much like social media platforms.

A second is the development of AI tools that are applicable to different populations. This calls for inclusive approaches to problem identification, incorporating perspectives from diverse communities and stakeholders. Public engagement, participatory research and transparent consultations are crucial for guiding the ethical and equitable development of AI in health care. These participatory approaches are discussed in section 3.2.8.

A third is access to advanced AI tools, such as those used in diagnostics or predictive analytics, in health facilities. Under-resourced settings may struggle to access or integrate these tools effectively, thereby exacerbating health inequities. At the same time, in places facing critical workforce shortages, AI could potentially improve access by filling care gaps; however, this also raises ethical concerns if only the privileged can consult with human health care professionals, while others are left to rely solely on AI. Low- and middle-income countries, in particular, risk being left behind in the AI health revolution, reinforcing the structural imbalance between the global north and south. The AI Preparedness Index, which includes macro-structural indicators relevant for AI adoption, such as digital infrastructure, innovation and economic integration, human capital

and regulation and ethics, shows how wealthier economies are better prepared than lower-income ones to adopt AI, even though there is significant variation between countries (Cazzaniga et al., 2024). The Government AI Readiness Index 2024, which examines 40 indicators across three pillars (Government, Technology Sector, and Data & Infrastructure), shows that North America and western European countries rank higher, although with some variation (Fig. 3.4), but the number of low- and middle-income countries developing AI national strategies is growing (Fuentes Nettel et al., 2024).

Fig. 3.4 *Government AI Readiness Index 2024 scores*



Source: Redrawn using data from Fuentes Nettel et al. (2024)

A fourth issue is digital exclusion. In increasingly digital societies, lack of Internet access risks excluding many people (van Deursen, 2020). This is often referred to as the digital divide and is shaped by factors such as age, gender and educational level (Lopez de Coca et al., 2022). In the United Kingdom, those most digitally excluded faced worse COVID-19 outcomes (Sounderajah et al., 2021). Policy responses, such as online schooling, further deepened inequalities as children without broadband or devices struggled to participate. In the USA, online searches for learning resources surged early in the pandemic, but mainly in wealthier, better-connected areas (Bacher-Hicks et al., 2021).

Digital-first health care models risk excluding vulnerable groups. These include online triage forms, appointment bookings, symptom checkers and telehealth services (McLean et al., 2013). Age is the main factor in digital exclusion, but disability, ethnicity, language, location, and low income also contribute (ONS, 2019). Groups at particular risk include refugees, asylum seekers, trafficked individuals, homeless people, sex workers, migrants with insecure status and Roma, Gypsy and Traveller communities (Doctors of the World, 2020).

Digital exclusion drives inequality. Those with low digital health literacy are more likely to experience poor health outcomes (NHS England, 2021). In the United Kingdom, 11.9 million people lack essential digital skills for accessing online health tools (Good Things Foundation, 2021).

The introduction of AI tools, particularly those interfacing directly with the public, could widen these gaps. First, disparities in digital literacy and digital exclusion may lead to uneven uptake of AI-enabled health services. Digital confidence is shaped by factors such as gender, income, age and digital access, and there is also a positive relationship between this confidence and attitudes towards AI (Bentley et al., 2024). Second, if these tools collect data for model improvement, the resulting feedback loops could skew AI systems further away from underrepresented groups, reinforcing their marginalization. Addressing digital barriers and understanding their impact on engagement with AI will be critical to preventing the deepening of digital exclusion and its negative consequences.

A fifth issue is the impact of AI on the job market. While AI holds promise for boosting productivity and automating administrative tasks, particularly through generative AI, its impact on employment remains a subject of ongoing debate. According to an International Monetary Fund study, nearly 40% of global jobs are exposed to AI (Cazzaniga et al., 2024). Importantly, while previous revolutions have tended to impact routine tasks through shifts towards automation, AI may instead also impact high-skilled jobs. The International Monetary Fund analysis finds that labour markets in advanced economies are more exposed than those in emerging markets, and within countries, disparities may widen as workers differ in their ability to leverage AI for productivity and income gains (Cazzaniga

et al., 2024). Although the future balance between automation and augmentation remains uncertain, it is evident that impacts will vary across countries, sectors and demographic groups (ILO, 2024).

Finally, as discussed by Morley et al. (2025), a significant concern in AI and health equity is the assumption that technological solutions alone, particularly those involving personalized medicine and advice, can meaningfully improve health outcomes. The narrative of AI-based personalized prevention may create the misleading impression that these are comprehensive preventive public health strategies. However, an exclusive focus on individual behaviour change, such as weight management through a chatbot, can divert attention and resources away from the broader determinants of health, including housing, income security and environmental conditions. This shift towards individual responsibility not only may reduce the role of the state in promoting healthy environments and risks increasing inequities, but also increases reliance on commercial apps and technologies that often lack public accountability.

In summary, in order to promote access and equality to AI systems, it is important to consider several dimensions across the AI development and deployment stages. This includes the sustained investment in digital infrastructure, particularly in underserved areas, to ensure that all communities, regardless of geography or socioeconomic status, can benefit from AI-enabled tools. It also involves digital and AI literacy across both the health workforce and the general population, including upskilling and reskilling initiatives for the current workforce. A failure to prioritize capacity-building and public education creates a real risk that AI tools will widen existing inequalities in access, trust and utilization. Inclusive public engagement is essential. Public health strategies should empower communities to guide the implementation of AI in their health systems, ensuring that the tools developed are not only technologically sound but also trusted and reflective of local realities. Governments could further promote equity by incentivizing innovation that addresses neglected diseases or the specific needs of underserved communities. Ultimately, policy-makers bear the responsibility not only to enable technological advancement but also to ensure that its benefits are socially just, inclusive and universally accessible.

3.2.8 How can we democratize the development of AI?

The emergence of participatory AI reflects a slowly emerging movement to involve diverse communities, especially historically marginalized groups, in the design, development and deployment of AI systems. Rooted in traditions of participatory design and democratic engagement, participatory AI recognizes that AI, while providing many opportunities for good, may also be a threat to social goods, justice and welfare (Buhmann & Fieseler, 2021; Cath, 2018) and it seeks

to shift power from centralized developers to the broader public. It emphasizes co-creation, transparency and empowerment, aiming to make AI more equitable and accountable. A notable example is a Citizens' Assembly undertaken by the Belgian government in 2024 (Box 3.11).

Box 3.11 *Citizen participation in shaping AI policy in Belgium*

In 2024, during its presidency of the Council of the EU, Belgium launched a notable initiative to involve ordinary citizens in shaping the future of AI policy (Permanent Representation of Belgium to the European Union, 2024). Sixty Belgian citizens, selected by lottery to represent the country's diversity, gathered over three weekends to discuss the societal, ethical and economic implications of AI. Their mission was to explore how AI should evolve within the EU and to offer guidance to policy-makers based on their collective vision. This initiative was not just a symbolic gesture. It was intended to demonstrate how public participation can enrich democratic governance, particularly in areas as complex and rapidly evolving as AI. The citizens, many of whom had little prior knowledge of AI, were given access to expert insights, hands-on demonstrations and structured discussions. Through this process, they developed a nuanced understanding of AI's potential and its risks, and they articulated a vision that emphasized responsibility, inclusivity and transparency.

One of the most striking, although perhaps unsurprising, outcomes of the panel was the identification of a significant knowledge gap between the public and the pace of AI development. Citizens expressed concern that they were largely uninformed about how AI is being developed and deployed, both in Belgium and across the EU. They called for large-scale public information campaigns and accessible educational resources to help bridge this gap. This call to action highlights the importance of transparency and effective communication in fostering public trust in emerging technologies.

The panel also highlighted the ethical dimensions of AI. Participants emphasized the importance of AI serving humanity, rather than replacing it. They stressed the importance of maintaining human oversight in critical areas such as health care, finance and defence. The principle of human-in-the-loop was seen as essential, not just at the final decision-making stage, but throughout the entire process. This insistence on human involvement reflects a deep concern for accountability, empathy and the preservation of human contact in an increasingly automated world.

Equity and inclusion were central themes throughout the panel's discussions. The citizens were adamant that no one should be left behind in the AI transition. They advocated for

> > *continues*

Box 3.11 *continued*

universal Internet access, equitable access to AI tools, and continuous education and retraining opportunities. They also called for protections for vulnerable groups and the preservation of creativity and soft skills, particularly in the face of automation. These priorities reflect a grounded understanding of the social fabric and the need to ensure that technological progress benefits all members of society.

Participation in the panel had a transformative effect on the citizens themselves. Surveys conducted before and after the panel showed a marked increase in trust in AI and in the belief that citizens can meaningfully contribute to complex policy issues. This demonstrates how participatory processes can empower individuals, foster civic engagement and improve the legitimacy of policy decisions.

The citizens' recommendations were both visionary and practical. They proposed integrating AI education into school curricula, developing eco-labels for AI systems, investing in European AI infrastructure and pursuing global ethical agreements. These proposals were grounded in deliberative consensus and offered a roadmap, both ambitious and socially attuned, that policy-makers can use.

The design of the panel itself was a model of inclusivity. Participants were selected through stratified random sampling to ensure diversity in age, gender, region, language and socioeconomic status. Barriers to participation were actively removed through measures such as childcare, travel support and multilingual facilitation. This inclusive approach not only enriched the deliberations but also demonstrated how democratic innovation can be made accessible to all.

The Belgian citizen panel on AI offers a thoughtful example of how public participation can shape the governance of emerging technologies. It shows that citizens are not only willing but capable of engaging with complex issues, and that their contributions can complement expert knowledge and enrich policy debates. As the EU advances in implementing the AI Act and navigating global AI competition, sustained and informed citizen involvement may be instrumental to ensuring that AI development aligns with democratic values and serves the public interest.

This case contributes to the growing body of evidence on the value and potential of deliberative democracy. It illustrates that involving citizens in policy-making is not just about consultation, it is about co-creation. In the context of AI, where decisions have far-reaching consequences, public participation is essential for building a future that is inclusive, ethical and human centred.

Source: Anne Swalue, International Relations Senior Expert, Federal Public Service Health, Food Chain Safety and Environment, Brussels, Belgium, personal communication, 2025.

A German team surveyed 3030 members of the general population on AI usage in patient health care, presenting them with hypothetical scenarios (vignettes) that varied in four key AI characteristics: autonomy, cost, reliability and transparency (Kuhne et al., 2025). Participants rated each scenario based on their support for AI use, perceived risk and expectations for personalized care. Most people had neutral to slightly negative attitudes towards AI in health care. Among the four attributes, reliability had the most decisive influence on public opinion. People wanted AI systems to be not only error-free but even more reliable than current methods. Transparency also mattered: people were wary of AI systems they could not understand or explain. In contrast, cost and autonomy had much less impact on attitudes. The authors also investigated whether factors such as age, gender, education or health status influenced opinions. While some differences emerged, such as education and migration background influencing views on transparency and autonomy, overall, sociodemographic differences were limited. The authors concluded that if the public is to accept AI in health care, systems must be reliable and transparent, and that concerns about cost or autonomy are secondary.

Birhane and colleagues (2022) have explored ways of making the development of AI tools more participatory. They argue that participatory methods can democratize AI development by centring the voices of marginalized communities, fostering inclusion and promoting justice, but that there is a lack of consensus on the precise meaning and goal of meaningful participation. They illustrate their arguments with three case studies. First, in a grass roots initiative, over 400 participants from over 20 countries collaborated to develop machine translation tools for African languages. Second, they describe how the Māori community in Aotearoa/New Zealand undertook a participatory project to record and annotate 300 hours of Te Reo Māori audio. This aimed to preserve and empower their language through the use of AI tools, such as speech recognition. Third, they described a participatory dataset documentation effort to improve dataset quality and validity, as a route to create responsible AI systems. However, the authors highlight several limitations of participatory approaches: participation can be coopted by powerful actors, reduced to token gestures or confused with mere inclusion or consultation. Without clear standards or accountability, it risks becoming a tool for legitimizing existing power structures rather than challenging them. The paper also cautions against using participation as a substitute for democratic governance and notes the challenges of accurately measuring its actual impact. Moreover, the burden placed on participants, especially marginalized communities, is often overlooked. Ultimately, the authors argue for a more reflexive, equitable, and context-sensitive approach to participatory AI, one that genuinely empowers communities rather than exploiting their involvement.

Chapter 4
Policy options

KEY MESSAGES

For governments

- Governments must establish robust, adaptable regulatory and evaluation frameworks to ensure AI systems in health care remain safe, effective and ethical throughout their life-cycles, with clear accountability and strong data protection.
- Strategic investment in AI infrastructure, workforce development and digital literacy is essential to build national readiness and support the secure, scalable deployment of AI across health systems.
- AI policy should prioritize equity, sustainability and human rights, embedding fairness and inclusion in design and governance while addressing environmental impacts and long-term societal implications.
- Global cooperation is critical to align national AI strategies with international norms and shared public health goals, ensuring responsible innovation that reflects diverse values and benefits all communities.

For health care providers and public health institutions

- Successful AI implementation in health care requires a structured, real-world approach, including piloting, monitoring and evaluation, supported by robust digital infrastructure, high-quality data and clear ethical guidelines.
- Organizations must invest in workforce readiness and strategic decision-making, balancing in-house development with external procurement, and ensuring staff are equipped through training and capacity-building to navigate the evolving AI landscape.
- Collaboration and equity should be central to AI adoption, with shared learning across organizations and inclusive design to ensure all populations benefit fairly from technological advances.

For the health professions

- Effective AI adoption in health care requires ongoing evaluation, informed decision-making and comprehensive workforce training, ensuring tools are safe, aligned with clinical needs and integrated responsibly into practice.
- Ethical and equitable use of AI must be prioritized, supported by interdisciplinary collaboration, strong governance and a commitment to serving all populations fairly across both health care and public health settings.

> > *continues*

KEY MESSAGES *continued*

For patients

- To ensure AI in health care serves all patients fairly and safely, policy must prioritize transparency, informed consent, inclusive design and meaningful patient participation, empowering individuals to understand, trust and shape how AI is used in their care.

4.1 Introduction

This chapter, drawing from the discussions in the previous chapters, outlines a range of policy options for governments, health care providers, public health institutions, professional bodies and patients to consider as they navigate the integration of AI into health systems and initiatives. It sets out a series of broad requirements but does not provide prescriptive recommendations on how to achieve them. Instead, it presents a menu of possible approaches, recognizing that the appropriate course of action will vary depending on national priorities, institutional capacities and local contexts.

The options presented reflect the complex and evolving nature of AI in health. They address key areas, including regulatory frameworks, data governance, workforce training, ethical oversight and environmental sustainability. Some focus on enabling innovation and collaboration, while others emphasize safeguards to ensure safety, equity and public trust, although clearly both are important. By providing a structured overview of these possibilities, this section aims to support informed decision-making and help policy-makers weigh the trade-offs as they develop their own strategies for the responsible adoption of AI.

Health professionals do not need to become AI experts. However, it is important that they develop a working understanding of the opportunities and the limitations of different types of AI models, such as the risks of bias in models and the potential for hallucinations in generative AI systems. This foundational knowledge is critical for making informed decisions about regulation, oversight and the responsible integration of AI into health systems and institutions.

4.2 For governments

4.2.1 System readiness and strategic integration

Governments must establish robust and adaptable regulatory and evaluation frameworks to ensure that AI systems in health care are safe, effective and ethical throughout their life-cycles. These frameworks should be agile enough to keep pace with innovation while maintaining rigorous standards for safety and efficacy. For countries operating within supranational regulatory systems or guidelines, such as the EU AI Act (or similar documents elsewhere, such as the guidance created by the Association of South East Asian Nations (ASEAN, 2024)), this may involve assessing the need for new national legislation, establishing dedicated AI governance bodies and ensuring the effective enforcement of these regulations. International instruments, including the Council of Europe's Framework Convention on AI, also shape regulatory approaches and foster cross-border cooperation. Regulatory sandboxes can provide a controlled environment for testing AI innovations before full-scale deployment, enabling the identification and mitigation of risks early on. Transparency throughout the regulatory process is essential to build trust among patients, clinicians and developers. While these regulatory frameworks are not without challenges, including the complexities of implementing risk-based approaches, they are indispensable for ensuring that AI technologies are deployed in ways that protect public health, uphold ethical standards and promote societal benefit. Effective evaluation of AI in health must integrate technical performance metrics with clinical, operational and societal considerations, requiring adapted technology assessment frameworks that address evolving challenges such as bias, explainability and system-level impact.

Governments must invest in secure, interoperable and scalable AI infrastructure to support the development of AI applications. A technically mature health AI ecosystem requires shared high-quality health data, secure data infrastructure, secure and scalable computing capacity, the ability to integrate into existing clinical workflows and MLOps capabilities. AI tools must be rigorously evaluated and approved by regulatory authorities before deployment and along their life-cycle. Data infrastructure must support continuous model updates and ensure secure, sustainable and equitable access to AI capabilities across regions and populations.

Building national AI readiness requires investment in digital literacy and capacity-building. Governments should assess their workforce's preparedness and implement upskilling and reskilling initiatives for health professionals. Policy-makers should develop relevant AI guidance and frameworks to guide health organizations in the adoption and implementation of AI solutions. Digital and AI literacy should also extend to the general population. Integrating technical

skills with critical thinking and socio-emotional competencies in education systems will bring benefits to society more generally in a world where AI will play a greater role in all aspects of life. Community engagement is vital to ensure that AI tools are trusted, contextually appropriate and aligned with local health needs. Without these efforts, AI risks exacerbating existing inequalities in access, trust and utilization.

Socioeconomic and environmental sustainability must be central to AI policy. While AI can improve operational efficiency and may support cost containment, it should complement, not replace, broader health system and social investments. At the heart of this is the recognition that AI technologies should be driven by clearly defined problems that stand to benefit from technological solutions. It is also important to acknowledge that many persistent health challenges stem from implementation barriers rather than a lack of innovation. To fully leverage the insights AI can offer, appropriate resources and mechanisms must be in place to translate these insights into meaningful action. Policy-makers should consider the full life-cycle costs of AI, including infrastructure, training, maintenance, computational running costs and environmental impact. Transparent reporting and international cooperation are essential to measure and mitigate AI's ecological footprint, ensuring that economic benefits do not come at the expense of environmental sustainability.

Global cooperation is essential to align national AI ambitions with international norms and ethical standards. Policy-makers must navigate the geopolitical dimensions of AI, striking a balance between domestic innovation and commitments to global equity and responsible governance. This includes promoting technology transfer, supporting equitable access to AI capabilities and learning from past global negotiations to address power imbalances and protect public interest. This will, however, be challenging in a situation in which the postwar multilateral trade architecture has been weakened (McKee et al., 2025a). Governments should be aware of the implications of global supply chain dependencies and the local impact of computing infrastructure, such as data centres.

4.2.2 Governance, ethics and rights

Ethical and human rights principles must underpin all AI health policy. Governments should ensure that AI systems are transparent, accountable and aligned with societal values and principles. This requires not only technical safeguards but also public dialogue about the ethical implications of increasingly autonomous systems. Ethical governance must evolve alongside technology to ensure that innovation enhances, rather than compromises, human dignity, equity and well-being.

Clear accountability, safety and human oversight are essential to building public trust in AI. Governments must establish liability frameworks that clearly define who is responsible when AI systems cause harm or fail to perform as intended. The principle of meaningful human control should guide all AI applications, particularly those with significant implications for health and safety. Human operators must be adequately trained to oversee AI systems technically and ethically, understanding the limitations and risks associated with different types of AI technologies. They should be designed with intuitive interfaces that support effective oversight. Safety must be embedded across the AI life-cycle, with continuous monitoring and the ability to deactivate or update systems as needed to protect public well-being.

Strong data protection and privacy safeguards are foundational for responsible AI development. Rather than hindering innovation, robust privacy policies enable the sustainable and ethical use of AI. Governments must strike a careful balance between enabling data-driven innovation and protecting individual rights. This includes ensuring informed consent, data literacy and ethical governance in both clinical and research contexts. As AI becomes increasingly present, particularly in public health, the tension between individual rights and collective benefits will require careful navigation. A particularly challenging issue, and one where regulatory approaches vary, relates to the intellectual property rights over training material.

The use of AI in health must be fair, equitable and inclusive. Policy-makers should consider fairness-aware machine learning practices and develop benchmarks to detect and mitigate the biases that can occur throughout the AI life-cycle. Legal frameworks must be updated or created to address discrimination in algorithmic decision-making. Equity-centred AI policies should include investments in digital infrastructure, especially in underserved areas and support the development of locally relevant AI models. Publicly funded AI initiatives should prioritize open-source development to prevent monopolization and provider capture and ensure public benefit. Inclusive design must also consider and reflect the needs of people with diverse backgrounds, prioritizing participatory approaches to ensure that AI tools are accessible and uphold human rights.

AI development and use should be decentralized and democratized through participatory approaches. Inclusive governance processes should involve diverse stakeholders, particularly those historically discriminated against in health systems and policies. Participatory design lowers barriers to entry and ensures that AI tools reflect the lived realities of different communities. Governments should promote decentralized and community-driven innovation to ensure that AI serves the public interest. Federated learning environments can support more equitable approaches to AI development and implementation, both within and between countries.

4.3 For health care providers and public health institutions

4.3.1 System readiness and strategic integration

AI implementation should follow a structured approach that includes piloting and monitoring and evaluating tools in real-world settings. Organizations must understand the relevant national regulations and guidance, as well as the limitations of monitoring and assessing procured AI products, especially when repurposing them for new applications. Establishing the necessary data pipelines and ensuring integration with existing systems are critical for effective deployment and oversight. Monitoring workforce trust and confidence in AI solutions will be important as part of the technology monitoring and evaluation process.

Health care organizations must ensure their workforce is equipped to navigate the evolving AI landscape. All staff involved in procuring, designing and operating AI tools should be familiar with relevant regulatory frameworks and national guidance, and possess the necessary skills to work effectively and ethically with AI technologies. This includes being aware of the human biases that can emerge when collaborating with AI in decision-making. This requires a dynamic and adaptive approach to workforce development, including continuous upskilling and the formation of multidisciplinary teams that include both digital and health expertise. Transparency in AI processes is essential for building trust among staff, patients and the community. Additionally, organizations should be mindful of potential skill erosion as AI automates certain tasks and take steps to preserve critical human expertise.

Organizations must carefully weigh the trade-offs between developing AI solutions in-house and procuring existing tools. Key considerations include cost, long-term dependencies, time to implementation, control over model biases and limitations, transparency, required staff expertise, datasets and the availability of computing and digital infrastructure. While developing local solutions may offer greater control and ability to customize to local needs, it also demands significant resources in terms of money, staff and time. Collaborating with third parties to co-develop tailored solutions can be a middle ground, although with careful attention to issues such as ownership of intellectual property. When developing or procuring tools, it is crucial to critically assess the risks and limitations of different AI technologies, ensure regulatory compliance and consider the hidden costs associated with using free, open-source models. Long-term factors such as scalability, flexibility and vendor lock-in should also be evaluated. Providers should ensure that AI tools have the relevant regulatory approvals.

Collaboration across organizations is essential to accelerate learning and maximize the benefits of AI. Sharing best practices, lessons learned and technical resources can help build a more cohesive and efficient AI ecosystem in health. Collaborative efforts can also reduce duplication and enhance scalability.

4.3.2 Governance, ethics and rights

The ethical use of AI must be promoted through clear guidelines that align with human rights and the public good. Engaging a broad range of stakeholders in discussions about AI's role will promote responsible use and equitable outcomes. Ethical frameworks should also uphold patient autonomy and informed consent, clearly communicating how AI will be used in care. Accountability mechanisms are essential to ensure safety, efficacy and public trust. Staff who will interact with AI systems that learn from feedback should be aware of the need for both technical and ethical oversight.

Robust digital infrastructure, high-quality data and strong governance systems are foundational to successful AI integration. Many institutions still rely on outdated information systems that may not be suited for AI applications. Secure environments for data processing are crucial, along with robust data governance and effective cybersecurity measures. This requires explicit and workable policies on data collection, storage and sharing. Cybersecurity protocols should be embedded in all operations (which is a principle that applies to any data processing operation, and not just those using AI). Clear guidelines on data ownership and access rights should empower patients and enhance trust in AI systems.

Equity must be a central principle in the design and implementation of AI in health care. To avoid reinforcing existing health disparities, training datasets must accurately reflect the diversity of the populations they serve. Prioritizing equity not only improves outcomes but also strengthens public trust in AI-driven health care.

4.4 For the health professions

4.4.1 System readiness and strategic integration

Ongoing evaluation and oversight are critical to ensuring AI tools are safe, effective and appropriate. Health professionals should engage with pilot projects and provide feedback on AI tools. They are often best placed to monitor outcomes, report concerns and understand the limitations of repurposing AI tools. This requires employers and professional bodies to ensure the necessary data pipelines and evaluation frameworks are in place to assess performance and impact.

Informed decision-making is essential when developing or procuring AI solutions. Health professionals must consider carefully when it is more appropriate to make or buy AI tools, taking account of hidden and long-term costs, technical dependencies, regulatory compliance and risks. They should ask questions about what their objective is, the extent of the evidence that it can be achieved, the sustainability of the tool and whether decisions are aligned with organizational goals and public health priorities. The purpose of AI and other technologies is to address real-world challenges. AI itself is a means to an end, not the end goal.

To fully harness the potential of AI, health professionals need comprehensive training in AI tools and their applications. Investing in upskilling programmes will help staff understand the strengths and limitations of AI, enabling them to use it responsibly and effectively. Training should cover core concepts in machine learning, data analytics and ethical considerations, while also fostering critical thinking to ensure that AI complements, rather than replaces, professional expertise. Staff should understand the different implications and risks associated with using AI to automate routine tasks versus using it to support decision-making. They should also be aware of the cognitive complexity in human–machine interaction, including the risks of automation bias. Accessible learning opportunities for all staff, from administrators to frontline workers, are essential for equitable adoption. It is important to remain mindful of skillset erosion as AI automates tasks, and ensure human expertise is maintained through continuous learning. Finally, there should be safeguards to protect whistleblowers who are concerned that AI tools are being implemented in ways that pose risks to patients.

Interdisciplinary collaboration and knowledge sharing are crucial to the successful adoption of AI. Success is more likely when designers and users of any technology collaborate meaningfully, but this is even more true with AI. This requires mechanisms to integrate perspectives from developers, ethicists, communities and diverse health care and public health professionals. These mechanisms should encourage the sharing of best practices and lessons learned across organizations to avoid duplication and accelerate responsible AI adoption.

4.4.2 Governance, ethics and rights

Ethical and equitable use of AI must be a priority in health care and public health. Health professionals should advocate for AI systems that are fair and trained on diverse datasets. They should understand the trade-offs between model complexity and explainability, and their implications for model transparency and human–machine interaction. They should also ensure that patients are informed about the role of AI in their care and that their autonomy is respected and maintained. They must be alert to potential biases and health disparities, promote inclusive

design and implementation practices and ensure compliance with relevant regulations and guidance before adopting AI tools in the workplace. Health professionals should understand data governance responsibilities, support secure environments for processing sensitive health information, ensure compliance with regulations such as the GDPR and advocate for systems that are resilient, scalable and interoperable.

4.5 For patients

As AI becomes increasingly embedded in health systems, patients are not only recipients of care but also active participants in shaping how these technologies are used, trusted and governed. While previous sections have focused on governments, health care providers and professionals, it is equally important to consider the role of patients and the policy options that support their rights, safety and meaningful engagement in an AI-enabled health system.

4.5.1 System readiness and strategic integration

Patients should be informed when AI is used in their care, including what data are being processed, how decisions are made and what safeguards are in place. This is particularly relevant in contexts where AI tools influence clinical decisions or generate health advice, such as virtual assistants or ambient AI scribes. Governments and health systems should invest in public-facing education initiatives to promote digital and AI literacy, tailored to different levels of understanding and cultural contexts. These efforts are essential to enable patients to critically assess AI-generated information and make informed choices about their care.

Application of AI systems in the health sector may rely on large volumes of personal health data, including sensitive information from electronic health records and diverse data sources. Patients must have confidence that their data are being used responsibly. Regulatory frameworks should be adapted to guarantee robust data protection, building on existing legislation such as secondary use of health data, including clear consent mechanisms, transparency about data use and the ability to opt out of certain applications. Special attention should be paid to the use of data from underserved populations, ensuring that AI systems do not reinforce existing inequities or compromise individual rights.

4.5.2 Governance, ethics and rights

AI tools must be designed and deployed in ways that serve all patients fairly. This includes ensuring that systems are trained on diverse datasets and tested across different populations. Patients from minority groups, those with disabilities and individuals in low-resource settings must be included in the design, evaluation

and governance of AI systems. Policy options should include requirements for inclusive design and equity audits, as well as mechanisms for patients to report concerns or adverse experiences with AI tools.

Patients should have a voice in shaping how AI is used in their health systems. Participatory governance models, such as citizens' panels, public consultations and co-design workshops, can help ensure that AI policies reflect the values and priorities of the communities they serve. These models should be inclusive, accessible and designed to support meaningful engagement. Governments and health institutions should establish mechanisms that enable patients to contribute to decisions regarding AI deployment, evaluation and oversight, including involvement in setting ethical standards and shaping the development of patient-facing tools.

Trust is essential for the successful integration of AI in health care. Patients must be confident that AI systems are safe, reliable and accountable. Policy frameworks should include clear mechanisms for oversight, including independent evaluation of AI tools, public reporting of performance metrics and pathways for redress when things go wrong. Patients should know who is responsible when AI systems fail or produce harmful outputs, and policies should ensure that affected individuals can seek justice and compensation.

Patients and communities are central to the future of AI in health. Their data powers many of these systems, their experiences shape the systems' effectiveness, and their trust determines the systems' success. Policy options must go beyond technical regulation to support patient rights, promote equity and enable meaningful participation. By embedding transparency, accountability and inclusion into AI governance, health systems can ensure that these technologies serve the public good and enhance, not undermine, human dignity in health care.

Chapter 5
Conclusion

Will AI solve the most pressing challenges in health, such as workforce shortages, rising costs, disparities in health outcomes and the difficulty of creating environments that support healthier choices? Will it empower individuals and communities to live healthier lives, or make our work more fulfilling by automating those tasks that are less enjoyable? Or will it do the opposite and displace jobs, deepen inequalities and further fragment our societies?

The answer, as with many transformative technologies, is that it depends. There is truth and misconception in all of these narratives. Some answers are already known, while others remain uncertain. What is clear, however, it is that the current wave of AI enthusiasm is accompanied by significant hype and myth-making. This can make it difficult for policy-makers, health care or public health organizations and health professionals, especially those new to the field, to make informed, critical decisions about how to engage with these technologies.

Urgent action is needed to ensure that AI is developed and deployed in a manner that serves the public interest. This means asking the right questions, grounded in principles of equity, transparency and accountability. It also means ensuring that the health sector, at all levels, is equipped to engage with AI not just as a tool, but as a socio-technical system with far-reaching implications. This engagement must be multiprofessional, drawing on the collective expertise of various disciplines to navigate the ethical, clinical, technical and societal dimensions of AI.

While those with extensive experience with AI may quickly grasp the implications of the more recent generative AI models, many health professionals are encountering these questions for the first time. This book aims to bridge that gap by providing a primer on what AI is, its potential impact on health and the policy options available to guide its responsible use.

AI is not new, but some of its capabilities are. Certain tasks will be fully automated, while others will require human oversight. The degree of automation appropriate in any given context will depend on factors such as the level of risk to human health, the nature of the task, the importance of human relationships and broader ethical, regulatory and feasibility considerations.

In some areas, AI may enhance efficiency and productivity. In others, it opens new frontiers for understanding health, such as identifying disease patterns or behavioural drivers. Yet, many of the most persistent health challenges are not due to a lack of technology or knowledge, but rather to implementation barriers, resource constraints and systemic issues that extend beyond the health sector.

To truly leverage AI's potential, we must recognize that technology alone is not a solution, but a means to an end. Without the proper infrastructure, safeguards and governance, AI risks exacerbating existing inequities and diverting resources from critical areas, such as social care or education.

The path forward requires thoughtful, inclusive and evidence-informed approaches. Regardless of where health professionals are in their journey of understanding AI and digital technologies, or the level of trust they place in them, it is essential to remain informed, balanced and critically engaged. Only by doing so can we ensure that AI serves as a force for equity, innovation and improved health outcomes, rather than a source of division or unintended harm.

Further reading

Here are some selected resources for those interested in further reading about AI, many of which are referenced throughout this book.

Resources on AI fundamentals and terminology

Council of Europe. **Artificial intelligence: glossary** [website]. Council of Europe. (https://www.coe.int/en/web/artificial-intelligence/glossary): the Council of Europe's glossary of AI terminology.

Goodfellow I et al. (2016). *Deep learning*. Cambridge, MA: The MIT Press. (https://www.deeplearningbook.org/): a book on the fundamentals of deep learning, available with free online access.

IBM. **What is artificial intelligence (AI)?** [website]. IBM. (https://www.ibm.com/think/topics/artificial-intelligence): IBM's *Think* platform has a wide range of short articles describing AI terminology some of which are referenced in this book.

ISO (2022). **ISO/IEC 22989:2022, Information technology – artificial intelligence – artificial intelligence concepts and terminology** [website]. International Organization for Standardization. (https://www.iso.org/standard/74296.html): a document that establishes terminology for AI and describes concepts in the field of AI.

POST (2024). **Artificial intelligence (AI) glossary** [website]. Parliamentary Office of Science and Technology, United Kingdom Parliament. (https://post.parliament.uk/artificial-intelligence-ai-glossary/): a glossary that compiles key terms used in recent Parliamentary Office of Science and Technology research on AI.

Russell S & Norvig P (2021). *Artificial intelligence: a modern approach*: a book covering a range of AI concepts and considerations for its use.

Resources on ethics and governance for the use of AI

OECD.AI Policy Observatory (2025). **OECD AI principles overview** [website]. OECD. (https://oecd.ai/en/ai-principles): the OECD AI Principles promote use of AI that is innovative and trustworthy and that respects human rights and democratic values, adopted in 2019 and updated in 2024.

WHO (2021). *Ethics and governance of artificial intelligence for health: WHO guidance*. Geneva: World Health Organization (https://iris.who.int/handle/10665/34199): a WHO report that identify the ethical challenges and risks on the use of AI for health and provides six consensus principles for its use.

WHO (2025). *Ethics and governance of artificial intelligence for health. Guidance on large multi-modal models*. Geneva: World Health Organization (https://www.who.int/publications/i/item/9789240084759): this WHO publication contains guidance specifically for using large multi-modal models (LMMs).

Landscape reports

AAAI (2025). *AAAI 2025 Presidential Panel on the future of AI research*. Washington, DC: Association for the Advancement of Artificial Intelligence (https://aaai.org/about-aaai/presidential-panel-on-the-future-of-ai-research/): this provides an overview of the AI research landscape.

Fuentes Nettel P et al. (2024). *Government AI Readiness Index 2024*. Great Malvern: Oxford Insights (https://oxfordinsights.com/ai-readiness/ai-readiness-index/): this is an annual report that focuses on the intersection of government and AI.

Maslej N et al. (2025). *Artificial intelligence index report 2025*. Stanford, CA: Institute for Human-Centered AI, Stanford University (https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf): Stanford University's Institute for Human-Centered AI annual report on the latest trends in the AI landscape.

Regulations governing the use of AI

Oviedo Convention – the only international legally binding instrument on the protection of human rights in the biomedical field

Council of Europe (1997). Convention for the protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine: Convention on Human Rights and Biomedicine (ETS No. 164) (https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/164).

General Data Protection Regulation – the European data privacy and security law

European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. Official Journal of the European Union. 119:1–88 (https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng).

Council of Europe's Framework Convention on Artificial Intelligence and Human Rights – the first-ever international legally binding treaty in AI and human right, democracy and the rule of law

Council of Europe (2024). Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225). (https://rm.coe.int/1680afb20c).

European Union Artificial Intelligence Act – a risk-based regulatory framework for AI systems in the European Union which has a phased implementation

European Union (2024). AI Act Regulation (EU) 2024/1689: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Publications Office of the European Union. (https://data.europa.eu/doi/10.2804/4225375).

References

AAAI (2025). AAAI 2025 Presidential Panel on the future of AI research. Washington, DC: Association for the Advancement of Artificial Intelligence (https://aaai.org/about-aaai/presidential-panel-on-the-future-of-ai-research/).

Abakasanga E et al. (2025). Equitable hospital length of stay prediction for patients with learning disabilities and multiple long-term conditions using machine learning. Front Digit Health.7:1538793. (https://doi.org/10.3389/fdgth.2025.1538793).

Abd-Alrazaq AA et al. (2019). An overview of the features of chatbots in mental health: a scoping review. Int J Med Inform.132:103978. (https://doi.org/10.1016/j.ijmedinf.2019.103978).

Aboy M et al. (2024). Navigating the EU AI Act: implications for regulated digital medical products. NPJ Digit Med.7(1):237. (https://doi.org/10.1038/s41746-024-01232-3).

Abramoff MD et al. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med.1(1):39. (https://doi.org/10.1038/s41746-018-0040-6).

Abramson J et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature.630(8016):493–500. (https://doi.org/10.1038/s41586-024-07487-w).

Adelani DI (2024). Meta's AI translation model embraces overlooked languages. Nature.630(8018):821–2. (https://doi.org/10.1038/d41586-024-00964-2).

Aggarwal A et al. (2023). Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. J Med Internet Res.25:e40789. (https://doi.org/10.2196/40789).

Agogo GO, Mwambi H (2025). Application of machine learning algorithms in an epidemiologic study of mortality. Ann Epidemiol.102:36–47. (https://doi.org/10.1016/j.annepidem.2024.12.015).

Ahmed N et al. (2023). The growing influence of industry in AI research. Science.379(6635):884–6. (https://doi.org/10.1126/science.ade2420).

Ahmed S et al. (2025). Can large language models challenge CNNs in medical image analysis? arXiv.2505:23503. (https://doi.org/10.48550/arxiv.2505.23503).

Ahmed Z et al. (2020). Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database (Oxford).2020:baaa010. (https://doi.org/10.1093/database/baaa010).

AI4HF (2025). Trustworthy artificial intelligence for personalised risk assessment in chronic heart failure (AI4HF) [website]. AI4HF. (https://www.ai4hf.com/about-ai4hf).

AIWaterUsage (2025). AIWaterUsage [website]. AIWaterUsage. (https://aiwaterusage.com/).

Ajanaku DF et al. (2025). California's AI laws are here—is your business ready. Pilbury (https://www.pillsburylaw.com/en/news-and-insights/california-ai-laws.html).

Akingbola A et al. (2024). Artificial intelligence and the dehumanization of patient care. J Med Surg Public Health.3:100138. (https://doi.org/10.1016/j.glmedi.2024.100138).

Al Meslamani AZ (2023). Beyond implementation: the long-term economic impact of AI in healthcare. J Med Econ.26(1):1566–9. (https://doi.org/10.1080/13696998.2023.2285186).

Alba C et al. (2025). The foundational capabilities of large language models in predicting postoperative risks using clinical notes. NPJ Digit Med.8(1):95. (https://doi.org/10.1038/s41746-025-01489-2).

Albrecht M et al. (2025). Enhancing clinical documentation with ambient artificial intelligence: a quality improvement survey assessing clinician perspectives on work burden, burnout, and job satisfaction. JAMIA Open.8(1):ooaf013. (https://doi.org/10.1093/jamiaopen/ooaf013).

Aldosery A et al. (2024). Enhancing public health response: a framework for topics and sentiment analysis of COVID-19 in the UK using Twitter and the embedded topic model. Front Public Health.12:1105383. (https://doi.org/10.3389/fpubh.2024.1105383).

ALGiT (2023). Exemplary case of work optimization with AI: Idrija Psychiatric Hospital [website]. ALGiT. (https://algit.si/en/ai-idrija-psychiatric-hospital/).

Ali S et al. (2023). Explainable Artificial Intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. Inf Fusion.99:101805. (https://doi.org/10.1016/j.inffus.2023.101805).

Allen B et al. (2024). PROVIDENT: development and validation of a machine learning model to predict neighborhood-level overdose risk in Rhode Island. Epidemiology.35(2):232–40. (https://doi.org/10.1097/EDE.0000000000001695).

Alshami A et al. (2023). Harnessing the power of ChatGPT for automating systematic review process: methodology, case study, limitations, and future directions. Systems.11(7):351. (https://doi.org/10.3390/systems11070351).

Amann J et al. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak.20(1):310. (https://doi.org/10.1186/s12911-020-01332-6).

Anthropic (2023). Claude's Constitution, 9 May 2023. San Francisco, CA: Anthropic (https://www.anthropic.com/news/claudes-constitution).

Anthropic (2025). Agentic misalignment: how LLMs could be insider threats. San Francisco, CA: Anthropic (https://www.anthropic.com/research/agentic-misalignment).

Anzolin G et al. (2024). Bridging the AI divide: empowering developing countries through manufacturing. Vienna: United Nations Industrial Development Organization (https://www.unido.org/sites/default/files/unido-publications/2024-07/IID%20Policy%20Brief%2012.pdf).

Armstrong S (2025). NHS England faces investigation over granting Foresight access to GP patient data. BMJ.389:r1192. (https://doi.org/10.1136/bmj.r1192).

Arnold KF et al. (2022). Estimating the effects of lockdown timing on COVID-19 cases and deaths in England: a counterfactual modelling study. PLoS One.17(4):e0263432. (https://doi.org/10.1371/journal.pone.0263432).

ASEAN (2024). ASEAN guide on AI governance and ethics. Jakarta: Association of Southeast Asian Nations (https://asean.org/wp-content/uploads/2024/02/ASEAN-Guide-on-AI-Governance-and-Ethics_beautified_201223_v2.pdf).

Asgari E et al. (2025). A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. npj Digit Med.8(1):274. (https://doi.org/10.1038/s41746-025-01670-7).

Aßmann E et al. (2025). Augmentation of wastewater-based epidemiology with machine learning to support global health surveillance. Nature Water.3(7):753–63. (https://doi.org/10.1038/s44221-025-00444-5).

Awad M, Khanna R (2015). Machine learning. In: Awad M, Khanna R, editors. Efficient learning machines. New York: Apress: 1–18.

Ayana G et al. (2024). Decolonizing global AI governance: assessment of the state of decolonized AI governance in Sub-Saharan Africa. R Soc Open Sci.11(8):231994. (https://doi.org/10.1098/rsos.231994).

Ayers JW et al. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med.183(6):589–96. (https://doi.org/10.1001/jamainternmed.2023.1838).

Bacher-Hicks A et al. (2021). Inequality in household adaptation to schooling shocks: Covid-induced online learning engagement in real time. J Public Econ.193:104345. (https://doi.org/10.1016/j.jpubeco.2020.104345).

Baclic O et al. (2020). Challenges and opportunities for public health made possible by advances in natural language processing. Can Commun Dis Rep.46(6):161–8. (https://doi.org/10.14745/ccdr.v46i06a02).

Bak M et al. (2022). You can't have AI both ways: balancing health data privacy and access fairly. Front Genet.13:929453. (https://doi.org/10.3389/fgene.2022.929453).

Bakken S (2023). AI in health: keeping the human in the loop. J Am Med Inform Assoc.30(7):1225–6. (https://doi.org/10.1093/jamia/ocad091).

Balloch J et al. (2024). Use of an ambient artificial intelligence tool to improve quality of clinical documentation. Future Healthc J.11(3):100157. (https://doi.org/10.1016/j.fhj.2024.100157).

Barrow N (2024). Anthropomorphism and AI hype. AI Ethic.4(3):707–11. (https://doi.org/10.1007/s43681-024-00454-1).

Bashingwa JJH et al. (2023). Can we design the next generation of digital health communication programs by leveraging the power of artificial intelligence to segment target audiences, bolster impact and deliver differentiated services? A machine learning analysis of survey data from rural India. BMJ Open.13(3):e063354. (https://doi.org/10.1136/bmjopen-2022-063354).

BBC (2023). Face search company Clearview AI overturns UK privacy fine. BBC. 18 October 2023 (https://www.bbc.co.uk/news/technology-67133157).

Beer D (2013). Algorithms: shaping tastes and manipulating the circulations of popular culture. In: Beer D, editor. Popular culture and new media: the politics of circulation: 63–100.

Bencevic M et al. (2024). Understanding skin color bias in deep learning-based skin lesion segmentation. Comput Methods Programs Biomed.245:108044. (https://doi.org/10.1016/j.cmpb.2024.108044).

Bentley SV et al. (2024). The digital divide in action: how experiences of digital technology shape future relationships with artificial intelligence. AI Ethics.4(4):901–15. (https://doi.org/10.1007/s43681-024-00452-3).

Bergmann D (2025). ELIZA effect at work: avoiding emotional attachment to AI coworkers [website]. IBM. (https://www.ibm.com/think/insights/eliza-effect-avoiding-emotional-attachment-to-ai).

Binesmael A et al. (2024). How does the public feel about health technologies and data?, London: The Health Foundation (https://www.health.org.uk/reports-and-analysis/analysis/how-does-the-public-feel-about-health-technologies-and-data).

Birhane A et al. (2022). Power to the people? Opportunities and challenges for participatory AI. EAAMO '22: Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, Arlington, VA.6:1–8. (https://doi.org/10.1145/3551624.3555290).

Birjandi SM, Khasteh SH (2021). A survey on data mining techniques used in medicine. J Diabetes Metab Disord.20(2):2055–71. (https://doi.org/10.1007/s40200-021-00884-2).

Blanco-Gonzalez A et al. (2023). The role of AI in drug discovery: challenges, opportunities, and strategies. Pharmaceuticals (Basel).16(6):891. (https://doi.org/10.3390/ph16060891).

Blili-Hamelin B et al. (2025). Stop treating 'AGI' as the north-star goal of AI research. arXiv.2502:03689. (https://doi.org/10.48550/arXiv.2502.03689).

BMA (2025). Use of Ambient scribes software in General Practice [website]. BMA. (https://cached.offlinehbpl.hbpl.co.uk/NewsAttachments/PGH/12105_ambient-scribes.pdf).

Boender TS et al. (2023). Establishing infodemic management in Germany: a framework for social listening and integrated analysis to report infodemic insights at the National Public Health Institute. JMIR Infodemiology.3:e43646. (https://doi.org/10.2196/43646).

Bojić L et al. (2024). Signs of consciousness in AI: can GPT-3 tell how smart it really is? Humanit Soc Sci.11(1):1631. (https://doi.org/10.1057/s41599-024-04154-3).

Bordukova M et al. (2024). Generative artificial intelligence empowers digital twins in drug discovery and clinical trials. Expert Opin Drug Discov.19(1):33–42. (https://doi.org/10.1080/17460441.2023.2273839).

Borowiec S (2016). AlphaGo seals 4-1 victory over Go grandmaster Lee Sedol. The Guardian. 15 March (https://www.theguardian.com/technology/2016/mar/15/googles-alphago-seals-4-1-victory-over-grandmaster-lee-sedol).

Bowe AK et al. (2023). Big data, machine learning, and population health: predicting cognitive outcomes in childhood. Pediatr Res.93(2):300–7. (https://doi.org/10.1038/s41390-022-02137-1).

Bowser DM et al. (2024). American clusters: using machine learning to understand health and health care disparities in the United States. Health Aff Sch.2(3):qxae017. (https://doi.org/10.1093/haschl/qxae017).

Britton A et al. (1999). Threats to applicability of randomised trials: exclusions and selective participation. J Health Serv Res Policy.4(2):112–21. (https://doi.org/10.1177/135581969900400210).

Brojka A et al. (2024). Governance of adaptive AI systems in medical devices. London: UCL Department of Science, Technology, Engineering and Public Policy (https://www.ucl.ac.uk/steapp/sites/steapp/files/8_final_report_-_governance_of_adaptive_ai_systems.pdf).

Brownstein JS et al. (2023). Advances in artificial intelligence for infectious-disease surveillance. N Engl J Med.388(17):1597–607. (https://doi.org/10.1056/NEJMra2119215).

Buhmann A, Fieseler C (2021). Towards a deliberative framework for responsible innovation in artificial intelligence. Technol Soc.64:101475. (https://doi.org/10.1016/j.techsoc.2020.101475).

Bui N et al. (2025). Fine-tuning large language models for improved health communication in low-resource languages. Comput Methods Programs Biomed.263:108655. (https://doi.org/10.1016/j.cmpb.2025.108655).

Buolamwini J, Gebru T (2018). Gender shades: intersectional accuracy disparities in commercial gender classification., Proceedings of Machine Learning Research: PMLR (https://proceedings.mlr.press/v81/buolamwini18a.html).

Cambria E et al. (2017). A practical guide to sentiment analysis. Cham: Springer.

Canadian Agency for Drugs and Technologies in Health (2024). Artificial intelligence for patient flow: emerging health technologies. Ottawa, ON: Canadian Agency for Drugs and Technologies in Health.

Cao Q et al. (2023). MepoGNN: metapopulation epidemic forecasting with graph neural networks. In: Amini M et al., editors. Machine learning and knowledge discovery in databases. Cham: Springer: 453–68.

Carciumaru TZ et al. (2025). Systematic review of machine learning applications using nonoptical motion tracking in surgery. NPJ Digit Med.8(1):28. (https://doi.org/10.1038/s41746-024-01412-1).

Cath C (2018). Governing artificial intelligence: ethical, legal and technical opportunities and challenges. Philos Trans R Soc A.376(2133):20180080. (https://doi.org/10.1098/rsta.2018.0080).

Cazzaniga M et al. (2024). Gen-AI: artificial intelligence and the future of work. Washington, DC: International Monetary Fund (https://www.elibrary.imf.org/view/journals/006/2024/001/006.2024.issue-001-en.xml).

CDC (2025). Artificial intelligence and machine learning: applying advanced tools for public health [website]. Centers for Disease Control. (https://www.cdc.gov/surveillance/data-modernization/technologies/ai-ml.html).

Challen R et al. (2019). Artificial intelligence, bias and clinical safety. BMJ Qual Saf.28(3):231–7. (https://doi.org/10.1136/bmjqs-2018-008370).

Chanda T et al. (2024). Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. Nat Commun.15(1):524. (https://doi.org/10.1038/s41467-023-43095-4).

Chen IY et al. (2020). Treating health disparities with artificial intelligence. Nat Med.26(1):16–7. (https://doi.org/10.1038/s41591-019-0649-2).

Chen J, See KC (2020). Artificial intelligence for COVID-19: rapid review. J Med Internet Res.22(10):e21476. (https://doi.org/10.2196/21476).

Chen M et al. (2024a). Opportunities and challenges of diffusion models for generative AI. Natl Sci Rev.11(12):nwae348. (https://doi.org/10.1093/nsr/nwae348).

Chen S (2025). How much energy will AI really consume? The good, the bad and the unknown. Nature.639(8053):22–4. (https://doi.org/10.1038/d41586-025-00616-z).

Chen S et al. (2024b). The effect of using a large language model to respond to patient messages. Lancet Digit Health.6(6):e379–e81. (https://doi.org/10.1016/S2589-7500(24)00060-8).

Chen Z (2023). Ethics and discrimination in artificial intelligence-enabled recruitment practices. Humanit Soc Sci.10(1):567. (https://doi.org/10.1057/s41599-023-02079-x).

Cheng M et al. (2024). ANTHROSCORE: a computational linguistic measure of anthropomorphism. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics.1:807–25. (https://doi.org/10.48550/arxiv.2402.02056).

Chi J et al. (2024). Artificial intelligence in metabolomics: a current review. Trends Analyt Chem.178:117852. (https://doi.org/10.1016/j.trac.2024.117852).

Chitty N, Dias S (2018). Artificial intelligence, soft power and social transformation. J Content Community Commun.7:1–14. (https://doi.org/10.31620/JCCC.06.18/02).

Choi WJ et al. (2024). A prediction of mutations in infectious viruses using artificial intelligence. Genomics Inform.22(1):15. (https://doi.org/10.1186/s44342-024-00019-y).

Chouffani El Fassi S et al. (2024). Not all AI health tools with regulatory authorization are clinically validated. Nat Med.30(10):2718–20. (https://doi.org/10.1038/s41591-024-03203-3).

Ciuti G et al. (2025). Robotic surgery. Nat Rev Bioeng.3(7):565–78. (https://doi.org/10.1038/s44222-025-00294-6).

Clapp S (2025). Defence and artificial intelligence. Brussels: European Parliamentary Research Service (https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2025)769580).

Clark J et al. (2025). Generative artificial intelligence use in evidence synthesis: a systematic review. Res Synth Methods.16(4):601–19. (https://doi.org/10.1017/rsm.2025.16).

Clark M, Bailey S (2024). Chatbots in health care: connecting patients to information. Can J Health Technol.4(1):1–21. (https://doi.org/https://doi.org/10.51731/cjht.2024.818).

Claudy MC et al. (2022). Artificial intelligence can't be charmed: the effects of impartiality on laypeople's algorithmic preferences. Front Psychol.13:898027. (https://doi.org/10.3389/fpsyg.2022.898027).

Cochrane (2025). (How) can AI-based automation tools assist with systematic searching? Artificial Intelligence (AI) methods in evidence synthesis: Cochrane Learning Live webinar series. London: Cochrane (https://training.cochrane.org/sites/training.cochrane.org/files/public/uploads/How%20can%20AI-based%20automation%20tools%20assist%20with%20systematic%20searching.pdf).

Cohen IG et al. (2023). How AI can learn from the law: putting humans in the loop only on appeal. NPJ Digit Med.6(1):160. (https://doi.org/10.1038/s41746-023-00906-8).

Colorado General Assembly (2024). Consumer protections for artificial intelligence. 2024 Regular Session. Denver, CO: Colorado General Assembly (https://leg.colorado.gov/bills/sb24-205).

Conroy G, Mallapaty S (2025). How China created AI model DeepSeek and shocked the world. Nature.638(8050):300–1. (https://doi.org/10.1038/d41586-025-00259-0).

Coser O et al. (2024). AI-based methodologies for exoskeleton-assisted rehabilitation of the lower limb: a review. Front Robot AI.11:1341580. (https://doi.org/10.3389/frobt.2024.1341580).

Council of Europe (1997). Oviedo Convention and its Protocols [website]. Council of Europe. (https://www.coe.int/en/web/human-rights-and-biomedicine/oviedo-convention).

Council of Europe (2024). Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. Strasbourg: Council of Europe (https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence).

Coupland H et al. (2025). Exploring the potential and limitations of deep learning and explainable AI for longitudinal life course analysis. BMC Public Health.25(1):1520. (https://doi.org/10.1186/s12889-025-22705-4).

Cruz Rivera S et al. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Lancet Digit Health.2(10):e549–e60. (https://doi.org/10.1016/S2589-7500(20)30219-3).

Cunard Chaney S, Nagi Mechael P (2020). Improving immunisation coverage and equity through the effectiveuse of geospatial technologies and data. Geneva: Gavi, the Vaccine Alliance (https://www.gavi.org/sites/default/files/document/2020/GIS-and-Immunisation-Landscape_EN.pdf).

Dahlgren G, Whitehead M (2021). The Dahlgren-Whitehead model of health determinants: 30 years on and still chasing rainbows. Public Health.199:20–4. (https://doi.org/10.1016/j.puhe.2021.08.009).

Darzi LA et al. (2018). Better health and care for all: a 10-point plan for the 2020s. London: Institute for Public Policy Research (https://www.ippr.org/articles/better-health-and-care-for-all).

De Luca S (2025). Algorithmic discrimination under the AI Act and the GDPR. Brussels: European Parliament Think Tank (https://www.europarl.europa.eu/thinktank/en/document/EPRS_ATA(2025)769509).

del Rey Guanter S (2024). El nuevo deber empresarial de "alfabetización tecnológica" de las personas trabajadoras sobre los sistemas de inteligencia artificial y la necesaria "alfabetización humanística" de estos sistemas. Madrid: AEDTSS (https://www.aedtss.com/el-nuevo-deber-empresarial-de-alfabetizacion-tecnologica-de-las-personas-trabajadoras-sobre-los-sistemas-de-inteligencia-artificial-y-la-necesaria-alfabetizacion-humanistica-de-estos-sistemas/).

del Rey Guanter S (2025). El difícil control jurídico de los sistemas avanzados de inteligencia artificial: un gran desafío para la Unión Europea y para las relaciones laborales. Madrid: AEDTSS (https://www.aedtss.com/el-dificil-control-juridico-de-los-sistemas-avanzados-de-inteligencia-artificial-un-gran-desafio-para-la-union-europea-y-para-las-relaciones-laborales/).

del Rey Puech P et al. (2025). Artificial intelligence and corruption: opportunities and challenges in the health sector. Int J Health Plann Mgmt. (https://doi.org/10.1002/hpm.70002).

Delord M et al. (2024). Patient-oriented unsupervised learning to uncover the patterns of multimorbidity associated with stroke using primary care electronic health records. BMC Prim Care.25(1):419. (https://doi.org/10.1186/s12875-024-02636-6).

Delpierre C, Lefevre T (2023). Precision and personalized medicine: what their current definition says and silences about the model of health they promote. Implication for the development of personalized health. Front Sociol.8:1112159. (https://doi.org/10.3389/fsoc.2023.1112159).

Dembrower K et al. (2023). Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. Lancet Digit Health.5(10):e703–e11. (https://doi.org/10.1016/S2589-7500(23)00153-X).

Deng L et al. (2023). Hospital crowdedness evaluation and in-hospital resource allocation based on image recognition technology. Sci Rep.13(1):299. (https://doi.org/10.1038/s41598-022-24221-6).

Department for Science, Innovation & Technology (2023). A pro-innovation approach to AI regulation [website]. United Kingdom Government. (https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper).

Di Bidino R et al. (2024). Health technology assessment framework for artificial intelligence-based technologies. Int J Technol Assess Health Care.40(1):e61. (https://doi.org/10.1017/S0266462324000308).

Disability Rights UK (2021). Self-driving cars pose threat for disabled people [website]. Disability Rights UK. (https://www.disabilityrightsuk.org/news/2021/april/self-driving-cars-pose-threat-disabled-people?srsltid=AfmBOoonXoEUKrtgAvF3OKQHLbzB0t1fRN48KSeOdfIpkkdtD-fVgoP0).

Discover-NOW (2020). The Discover Dataset [website]. London Secure Data Environment. (https://discover-now.co.uk/the-data/).

Doctors of the World (2020). A rapid needs assessment of excluded people in England during the 2020 COVID-19 pandemic. London: Doctors of the World (http://www.doctorsoftheworld.org.uk/wp-content/uploads/2020/05/covid19-full-rna-report.pdf).

Doll R, Hill AB (1956). Lung cancer and other causes of death in relation to smoking; a second report on the mortality of British doctors. Br Med J.2(5001):1071–81. (https://doi.org/10.1136/bmj.2.5001.1071).

Duggan MJ et al. (2025). Clinician experiences with Ambient Scribe technology to assist with documentation burden and efficiency. JAMA Netw Open.8(2):e2460637. (https://doi.org/10.1001/jamanetworkopen.2024.60637).

Dugger SA et al. (2018). Drug development in the era of precision medicine. Nat Rev Drug Discov.17(3):183–96. (https://doi.org/10.1038/nrd.2017.226).

Duong D, Solomon BD (2025). Artificial intelligence in clinical genetics. Eur J Hum Genet.33(3):281–8. (https://doi.org/10.1038/s41431-024-01782-w).

Durán JM, Pozzi G (2025). Trust and trustworthiness in AI. Philos Technol.38(1). (https://doi.org/10.1007/s13347-025-00843-2).

Dyar OJ et al. (2022). Rainbows over the world's public health: determinants of health models in the past, present, and future. Scand J Public Health.50(7):1047–58. (https://doi.org/10.1177/14034948221113147).

Efthimiou O et al. (2024). Developing clinical prediction models: a step-by-step guide. BMJ.386:e078276. (https://doi.org/10.1136/bmj-2023-078276).

Epic (2025). AI for operations [website]. Epic. (https://www.epic.com/software/ai-operations/).

Epley N et al. (2007). On seeing human: a three-factor theory of anthropomorphism. Psychol Rev.114(4):864–86. (https://doi.org/10.1037/0033-295X.114.4.864).

Es S et al. (2025). Ragas: automated evaluation of retrieval augmented generation. arXiv.2309.15217. (https://doi.org/10.48550/arXiv.2309.15217).

Esteva A et al. (2021). Deep learning-enabled medical computer vision. NPJ Digit Med.4(1):5. (https://doi.org/10.1038/s41746-020-00376-2).

Esteva A et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature.542(7639):115–8. (https://doi.org/10.1038/nature21056).

Esteva A et al. (2019). A guide to deep learning in healthcare. Nat Med.25(1):24–9. (https://doi.org/10.1038/s41591-018-0316-z).

EuroLLM team et al. (2024). EuroLLM-9B. New York: Hugging Face (https://huggingface.co/blog/eurollm-team/eurollm-9b).

European Commission (2019). Ethics guidelines for trustworthy artificial intelligence. Brussels: European Commission (https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai).

European Commission (2021). Coordinated plan on artificial intelligence 2021 review. Brussels: European Commission (https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review).

European Commission (2023a). EU-U.S. terminology and taxonomy for artificial intelligence. Brussels: European Commission (https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence).

European Commission (2023b). ?A European Health Data Space: Questions and Answers

European Commission (2024). Commission establishes AI Office to strengthen EU leadership in safe and trustworthy Artificial Intelligence [press release]. 29 May. Brussels: European Commission (https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2982).

European Commission (2025a). Artificial intelligence in healthcare [website]. European Commission. (https://health.ec.europa.eu/ehealth-digital-health-and-care/artificial-intelligence-healthcare_en).

European Commission (2025b). New EU rules on Health Technology Assessment open up a new era for patient access to innovation [press release]. 10 Jan. Brussels: European Commission (https://ec.europa.eu/commission/presscorner/detail/en/ip_25_226).

European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. Brussels: European Union (https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng).

European Union (2017). Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices. Brussels: European Union (https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng).

European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Brussels: European Union (https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689).

Fairley L et al. (2011). The influence of both individual and area based socioeconomic status on temporal trends in Caesarean sections in Scotland 1980–2000. BMC Public Health.11:330. (https://doi.org/10.1186/1471-2458-11-330).

Farooq K, Solowiej BJ (2021). Artificial intelligence in the public sector: maximizing opportunities, managing risks. Washington, DC: World Bank Group (http://documents.worldbank.org/curated/en/809611616042736565).

Feigerlova E et al. (2025). A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. BMC Med Educ.25(1):129. (https://doi.org/10.1186/s12909-025-06719-5).

Feng H et al. (2024). An adaptive decision support system for outpatient appointment scheduling with heterogeneous service times. Sci Rep.14(1):27731. (https://doi.org/10.1038/s41598-024-77873-x).

Ferguson T et al. (2022). Effectiveness of wearable activity trackers to increase physical activity and improve health: a systematic review of systematic reviews and meta-analyses. Lancet Digit Health.4(8):e615–e26. (https://doi.org/10.1016/S2589-7500(22)00111-X).

Feuerriegel S et al. (2024). Causal machine learning for predicting treatment outcomes. Nat Med.30(4):958–68. (https://doi.org/10.1038/s41591-024-02902-1).

Filippi E et al. (2023). Automation technologies and their impact on employment: a review, synthesis and future research agenda. Technol Forecast Soc Change.191:122448. (https://doi.org/10.1016/j.techfore.2023.122448).

Fisher S, Rosella LC (2022). Priorities for successful use of artificial intelligence by public health organizations: a literature review. BMC Public Health.22(1):2146. (https://doi.org/10.1186/s12889-022-14422-z).

Folk DP et al. (2025). Cultural variation in attitudes toward social chatbots. J Cross Cult Psychol.56(3):219–39. (https://doi.org/10.1177/00220221251317950).

Fontes C et al. (2022). AI-powered public surveillance systems: why we (might) need them and how we want them. Technol Soc.71:102137. (https://doi.org/10.1016/j.techsoc.2022.102137).

Foretz M et al. (2023). Metformin: update on mechanisms of action and repurposing potential. Nat Rev Endocrinol.19(8):460–76. (https://doi.org/10.1038/s41574-023-00833-4).

Fountzilas E et al. (2025). Convergence of evolving artificial intelligence and machine learning techniques in precision oncology. NPJ Digit Med.8(1):75. (https://doi.org/10.1038/s41746-025-01471-y).

Fraser H et al. (2018). Safety of patient-facing digital symptom checkers. Lancet.392(10161):2263–4. (https://doi.org/10.1016/S0140-6736(18)32819-8).

Fu T et al. (2022). HINT: hierarchical interaction network for clinical-trial-outcome predictions. Patterns (N Y).3(4):100445. (https://doi.org/10.1016/j.patter.2022.100445).

Fuentes Nettel P et al. (2024). Government Al Readiness Index 2024. Great Malvern: Oxford Insights (https://oxfordinsights.com/ai-readiness/ai-readiness-index/).

Furlan R et al. (2021). A natural language processing-based virtual patient simulator and intelligent tutoring system for the clinical diagnostic process: simulator development and case study. JMIR Med Inform.9(4):e24073. (https://doi.org/10.2196/24073).

Fuster-Casanovas A et al. (2025). Evaluating patient and professional satisfaction and documentation time reduction through an AI-driven automatic clinical note generation in primary care: a proof of concept (preprint). JMIR Preprints.29/07/2025:80549. (https://doi.org/10.2196/preprints.80549).

Gallagher K et al. (2024). Mathematical model-driven deep learning enables personalized adaptive therapy. Cancer Res.84(11):1929–41. (https://doi.org/10.1158/0008-5472.CAN-23-2040).

Gallo V, Nair S (2024). The UK's framework for AI regulation. London: Deloitte (https://www.deloitte.com/uk/en/Industries/financial-services/blogs/the-uks-framework-for-ai-regulation.html).

Garcia D (2023). Algorithms and decision-making in military artificial intelligence. Global Society.38(1):24–33. (https://doi.org/10.1080/13600826.2023.2273484).

GBD 2021 Forecasting Collaborators (2024). Burden of disease scenarios for 204 countries and territories, 2022–2050: a forecasting analysis for the Global Burden of Disease Study 2021. Lancet.403(10440):2204–56. (https://doi.org/10.1016/S0140-6736(24)00685-8).

Gianfrancesco MA et al. (2018). Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med.178(11):1544–7. (https://doi.org/10.1001/jamainternmed.2018.3763).

Gichoya JW et al. (2022). AI recognition of patient race in medical imaging: a modelling study. Lancet Digit Health.4(6):e406–e14. (https://doi.org/10.1016/S2589-7500(22)00063-2).

Gilbert S et al. (2020). How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs. BMJ Open.10(12):e040269. (https://doi.org/10.1136/bmjopen-2020-040269).

Goddard K et al. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc.19(1):121–7. (https://doi.org/10.1136/amiajnl-2011-000089).

Goh E et al. (2024). Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open.7(10):e2440969. (https://doi.org/10.1001/jamanetworkopen.2024.40969).

Goh KH et al. (2021). Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nat Commun.12(1):711. (https://doi.org/10.1038/s41467-021-20910-4).

Goldman AI (1979). What is justified belief? In: Pappas GS, editor. Justification and knowledge: new studies in epistemology. Dordrecht: Springer: 1–23.

Goldstein BA et al. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc.24(1):198–208. (https://doi.org/10.1093/jamia/ocw042).

Gomez LRE et al. (2025). Holistic forecasting for future pandemics: a review of pathogens, models, and data. Discover Public Health.22(1):211. (https://doi.org/10.1186/s12982-025-00573-y).

Good Things Foundation (2021). Health and wellbeing [website]. Good Things Foundation. (https://www.goodthingsfoundation.org/areas-of-work/health-and-wellbeing).

Goodfellow I et al. (2016). Deep learning. Cambridge, MA: The MIT Press.

Goodfellow I et al. (2020). Generative adversarial networks. Communications of the ACM.63(11):139–44. (https://doi.org/10.1145/3422622).

Goodfellow IJ et al. (2015). Explaining and harnessing adversarial examples. arXiv.1412:6572. (https://doi.org/10.48550/arXiv.1412.6572).

Gordon M et al. (2024). A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach.46(4):446–70. (https://doi.org/10.1080/0142159X.2024.2314198).

Grah JS et al. (2025). [Applications, challenges and a trustworthy use of artificial intelligence in public health]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz.68(8):880–8. (https://doi.org/10.1007/s00103-025-04098-2).

Greatbatch D et al. (2005). Telephone triage, expert systems and clinical expertise. Sociol Health Illn.27(6):802–30. (https://doi.org/10.1111/j.1467-9566.2005.00475.x).

Greenhalgh T et al. (2017). Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res.19(11):e367. (https://doi.org/10.2196/jmir.8775).

Guan Y et al. (2025). Keeping medical AI healthy: a review of detection and correction methods for system degradation. arXiv.2506.17442. (https://doi.org/10.48550/arXiv.2506.17442).

Guevara M et al. (2024). Large language models to identify social determinants of health in electronic health records. NPJ Digit Med.7(1):6. (https://doi.org/10.1038/s41746-023-00970-0).

Guo LL et al. (2023). EHR foundation models improve robustness in the presence of temporal distribution shift. Sci Rep.13(1):3767. (https://doi.org/10.1038/s41598-023-30820-8).

Habicht J et al. (2024). Closing the accessibility gap to mental health treatment with a personalized self-referral chatbot. Nat Med.30(2):595–602. (https://doi.org/10.1038/s41591-023-02766-x).

Haque A et al. (2020). Illuminating the dark spaces of healthcare with ambient intelligence. Nature.585(7824):193–202. (https://doi.org/10.1038/s41586-020-2669-y).

Haque MDR, Rubya S (2023). An overview of chatbot-based mobile mental health apps: insights from app description and user reviews. JMIR Mhealth Uhealth.11:e44838. (https://doi.org/10.2196/44838).

Hasanzadeh F et al. (2025). Bias recognition and mitigation strategies in artificial intelligence healthcare applications. NPJ Digit Med.8(1):154. (https://doi.org/10.1038/s41746-025-01503-7).

Hassan N et al. (2024). Artificial intelligence informing clinical decision making on the risk of hospital readmissions in multi-morbid patients: a systematic review. Int J Pharm Pract.32(Supplement_1):i40–i1. (https://doi.org/10.1093/ijpp/riae013.050).

Hicks MT et al. (2024). ChatGPT is bullshit. Ethics Inform Technol.26(2):1. (https://doi.org/10.1007/s10676-024-09775-5).

Hill AB (1965). The environment and disease: association or causation? Proc Royal Soc Med.58(5):295–300. (https://doi.org/10.1177/003591576505800503).

Hodson TO (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geosci Model Dev.15(14):5481–7. (https://doi.org/10.5194/gmd-15-5481-2022).

Holm NN et al. (2025). amVAE: Age-aware multimorbidity clustering using variational autoencoders. Comput Biol Med.186:109632. (https://doi.org/10.1016/j.compbiomed.2024.109632).

Hotez PJ et al. (2020). Combating vaccine hesitancy and other 21st century social determinants in the global fight against measles. Curr Opin Virol.41:1–7. (https://doi.org/10.1016/j.coviro.2020.01.001).

Hou L et al. (2024). An autonomous wheelchair with health monitoring system based on Internet of Thing. Sci Rep.14(1):5878. (https://doi.org/10.1038/s41598-024-56357-y).

Huang D et al. (2024). From large language models to large multimodal models: a literature review. Applied Sciences.14(12):5068. (https://doi.org/10.3390/app14125068).

Hughes L et al. (2025). AI agents and agentic systems: a multi-expert analysis. J Comput Inf Systems.65(4):489–517. (https://doi.org/10.1080/08874417.2025.2483832).

Huo B et al. (2025). Large language models for chatbot health advice studies: a systematic review. JAMA Netw Open.8(2):e2457879. (https://doi.org/10.1001/jamanetworkopen.2024.57879).

Hurley ME et al. (2025). Patient consent and the right to notice and explanation of AI systems used in health care. Am J Bioeth.25(3):102–14. (https://doi.org/10.1080/15265161.2024.2399828).

Hutson M (2024). How AI is being used to accelerate clinical trials. Nature.627(8003):S2–S5. (https://doi.org/10.1038/d41586-024-00753-x).

Hwang M et al. (2025). AI applications for chronic condition self-management: scoping review. J Med Internet Res.27:e59632. (https://doi.org/10.2196/59632).

IBM (2023a). What are AI hallucinations? [website]. IBM. (https://www.ibm.com/think/topics/ai-hallucinations).

IBM (2023b). What is retrieval-augmented generation? [website]. IBM. (https://research.ibm.com/blog/retrieval-augmented-generation-RAG).

IBM (2024a). Use gen AI economics to lap the competition [website]. IBM's Institute for Business Value. (https://www.ibm.com/thought-leadership/institute-business-value/report/ceo-generative-ai/ceo-ai-cost-of-compute).

IBM (2024b). What are diffusion models? [website]. IBM. (https://www.ibm.com/think/topics/diffusion-models).

IBM (2024c). What is fine-tuning? [website]. IBM. (https://www.ibm.com/think/topics/fine-tuning).

IBM (2025a). What are AI agents? [website]. IBM. (https://www.ibm.com/think/topics/ai-agents).

IBM (2025b). What is edge AI? [website]. IBM. (https://www.ibm.com/think/topics/edge-ai).

Ibrahim M et al. (2025). Generative AI for synthetic data across multiple medical modalities: a systematic review of recent developments and challenges. Comput Biol Med.189:109834. (https://doi.org/10.1016/j.compbiomed.2025.109834).

ICO (2023). Information Commissioner seeks permission to appeal Clearview AI Inc ruling. Information Commissioner's Office. 17 November (https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2023/11/information-commissioner-seeks-permission-to-appeal-clearview-ai-inc-ruling/).

ICO (2025a). AI and data protection risk toolkit [website]. UIK Government. (https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/ai-and-data-protection-risk-toolkit/).

ICO (2025b). UK Upper Tribunal hands down judgment on Clearview AI Inc. Information Commissioner's Office. 8 October (https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2025/10/uk-upper-tribunal-hands-down-judgment-on-clearview-ai-inc/).

IEEE (2024). Do we dare use generative AI for mental health? [website]. IEEE. (https://spectrum.ieee.org/woebot).

Iftikhar M et al. (2024). Artificial intelligence: revolutionizing robotic surgery: review. Ann Med Surg (Lond).86(9):5401–9. (https://doi.org/10.1097/MS9.0000000000002426).

ILO (2024). What is the possible effect of generative AI on employment? [website]. ILO. (https://www.ilo.org/resource/other/what-possible-effect-generative-ai-employment).

ISO (2020a). Clinical investigation of medical devices for human subjects — good clinical practice. Geneva: International Organization for Standardization (https://www.iso.org/standard/71690.html).

ISO (2020b). Medical devices — Post-market surveillance for manufacturers. Geneva: International Organization for Standardization (https://www.iso.org/standard/67942.html).

Jackson GP, Shortliffe EH (2025). Understanding the evidence for artificial intelligence in healthcare. BMJ Qual Saf.34(7):421–4. (https://doi.org/10.1136/bmjqs-2025-018559).

Jalil Z et al. (2021). COVID-19 related sentiment analysis using state-of-the-art machine learning and deep learning techniques. Front Public Health.9:812735. (https://doi.org/10.3389/fpubh.2021.812735).

James N (2024). Access to services: the promises and pitfalls of AI for people with disabilities. European Disability Forum. 13 January (https://www.edf-feph.org/access-to-services-the-promises-and-pitfalls-of-ai-for-people-with-disabilities/).

Janiesch C et al. (2021). Machine learning and deep learning. Electron Mark.31(3):685–95. (https://doi.org/10.1007/s12525-021-00475-2).

Jayaraman P et al. (2024). A primer on reinforcement learning in medicine for clinicians. NPJ Digit Med.7(1):337. (https://doi.org/10.1038/s41746-024-01316-0).

Jin Q et al. (2024). Matching patients to clinical trials with large language models. Nat Commun.15(1):9074. (https://doi.org/10.1038/s41467-024-53081-z).

Johnson KB et al. (2021). Precision medicine, AI, and the future of personalized health care. Clin Transl Sci.14(1):86–93. (https://doi.org/10.1111/cts.12884).

Jones N (2025). AI hallucinations can't be stopped - but these techniques can limit their damage. Nature.637(8047):778–80. (https://doi.org/10.1038/d41586-025-00068-5).

Joshi A et al. (2020). Harnessing tweets for early detection of an acute disease event. Epidemiology.31(1):90–7. (https://doi.org/10.1097/EDE.0000000000001133).

Kak A, West SM (2023). AI Now 2023 landscape: confronting tech power. New York: AI Now Institute (https://ainowinstitute.org/2023-landscape).

Kalluri PR et al. (2025). Computer-vision research powers surveillance technology. Nature.643(8070):73–9. (https://doi.org/10.1038/s41586-025-08972-6).

Kannel WB et al. (1961). Factors of risk in the development of coronary heart disease – six year follow-up experience. The Framingham Study. Ann Intern Med.55:33–50. (https://doi.org/10.7326/0003-4819-55-1-33).

Kaplan AD et al. (2023). Trust in artificial intelligence: meta-analytic findings. Hum Factors.65(2):337–59. (https://doi.org/10.1177/00187208211013988).

Kastrup N et al. (2024). Landscape and challenges in economic evaluations of artificial intelligence in healthcare: a systematic review of methodology. BMC Digital Health.2(1):39. (https://doi.org/10.1186/s44247-024-00088-7).

Katsoulakis E et al. (2024). Digital twins for health: a scoping review. NPJ Digit Med.7(1):77. (https://doi.org/10.1038/s41746-024-01073-0).

Kavalci E, Hartshorn A (2023). Improving clinical trial design using interpretable machine learning based prediction of early trial termination. Sci Rep.13(1):121. (https://doi.org/10.1038/s41598-023-27416-7).

Kelly CJ et al. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Med.17(1):195. (https://doi.org/10.1186/s12916-019-1426-2).

Khalifa M, Albadawy M (2024). AI in diagnostic imaging: revolutionising accuracy and efficiency. Comp Methods Programs Biomed Update.5:100146. (https://doi.org/10.1016/j.cmpbup.2024.100146).

Kim E et al. (2023). Factors affecting success of new drug clinical trials. Ther Innov Regul Sci.57(4):737–50. (https://doi.org/10.1007/s43441-023-00509-1).

Kim J et al. (2025). Artificial intelligence tools in supporting healthcare professionals for tailored patient care. NPJ Digit Med.8(1):210. (https://doi.org/10.1038/s41746-025-01604-3).

King Z et al. (2022). Machine learning for real-time aggregated prediction of hospital admission for emergency patients. NPJ Digit Med.5(1):104. (https://doi.org/10.1038/s41746-022-00649-y).

Knudsen JE et al. (2024). Clinical applications of artificial intelligence in robotic surgery. J Robot Surg.18(1):102. (https://doi.org/10.1007/s11701-024-01867-0).

Köbis N et al. (2022). The promise and perils of using artificial intelligence to fight corruption. Nat Mach Intell.4(5):418–24. (https://doi.org/10.1038/s42256-022-00489-1).

Kocuvan P et al. (2024). Enhancing healthcare with intelligent environments: Integrating medical knowledge into GPT for advanced medical personal chatbots. J Smart Cities Soc.3(3):177–92. (https://doi.org/10.3233/scs-240011).

Kokosi T, Harron K (2022). Synthetic data in medical research. BMJ Med.1(1):e000167. (https://doi.org/10.1136/bmjmed-2022-000167).

Kosmyna N et al. (2025). Your brain on ChatGPT: accumulation of cognitive debt when using an AI assistant for essay writing task. arXiv.2506:08872. (https://doi.org/10.48550/arxiv.2506.08872).

Kponyo JJ et al. (2024). Techno-neocolonialism: an emerging risk in the artificial intelligence revolution. Trayectorias Humanas Trascontinentales.18 (18). (https://doi.org/10.25965/trahs.6382).

Kraemer MUG et al. (2025). Artificial intelligence for modelling infectious disease epidemics. Nature.638(8051):623–35. (https://doi.org/10.1038/s41586-024-08564-w).

Kreuzberger D et al. (2022). Machine learning operations (MLOps): overview, definition, and architecture. CoRR.abs/2205.02302. (https://doi.org/10.48550/ARXIV.2205.02302).

Krieger M et al. (2025). An overdose forecasting dashboard for local harm-reduction response. Health Promot Pract.15248399251335620. (https://doi.org/10.1177/15248399251335620).

Kringos D et al. (2013). The strength of primary care in Europe: an international comparative study. Br J Gen Pract.63(616):e742–50. (https://doi.org/10.3399/bjgp13X674422).

Kristensen FB et al. (2017). The HTA Core Model((R)) - 10 years of developing an international framework to share multidimensional value assessment. Value Health.20(2):244–50. (https://doi.org/10.1016/j.jval.2016.12.010).

Kuhne S et al. (2025). Attitudes toward AI usage in patient health care: evidence from a population survey vignette experiment. J Med Internet Res.27:e70179. (https://doi.org/10.2196/70179).

Kumazu Y et al. (2021). Automated segmentation by deep learning of loose connective tissue fibers to define safe dissection planes in robot-assisted gastrectomy. Sci Rep.11(1):21198. (https://doi.org/10.1038/s41598-021-00557-3).

Lagemann K et al. (2023). Deep learning of causal structures in high dimensions under data limitations. Nat Mach Intell.5(11):1306–16. (https://doi.org/10.1038/s42256-023-00744-z).

Lancet (2025). Health in the age of disinformation. Lancet.405(10474):173. (https://doi.org/10.1016/S0140-6736(25)00094-7).

Laranjo L et al. (2018). Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc.25(9):1248–58. (https://doi.org/10.1093/jamia/ocy072).

Laurence T et al. (2025). Review GIDE – restaurant review gastrointestinal illness detection and extraction with large language models. arXiv.2503.09743. (https://doi.org/10.48550/arXiv.2503.09743).

Laymouna M et al. (2024). Roles, users, benefits, and limitations of chatbots in health care: rapid review. J Med Internet Res.26:e56930. (https://doi.org/10.2196/56930).

Leslie D (2019). Understanding artificial intelligence ethics and safety: a guide for the responsible design and implementation of AI systems in the public sector. London: The Alan Turing Institute (https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf).

Li H et al. (2023a). Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med.6(1):236. (https://doi.org/10.1038/s41746-023-00979-5).

Li X et al. (2023b). Machine learning methods for accurately predicting survival and guiding treatment in stage I and II hepatocellular carcinoma. Medicine (Baltimore).102(45):e35892. (https://doi.org/10.1097/MD.0000000000035892).

Lieberum JL et al. (2025). Large language models for conducting systematic reviews: on the rise, but not yet ready for use - a scoping review. J Clin Epidemiol.181:111746. (https://doi.org/10.1016/j.jclinepi.2025.111746).

Liefgreen A et al. (2024). Beyond ideals: why the (medical) AI industry needs to motivate behavioural change in line with fairness and transparency values, and how it can do it. AI Soc.39(5):2183–99. (https://doi.org/10.1007/s00146-023-01684-3).

Lim S, Schmälzle R (2023). Artificial intelligence for health message generation: an empirical study using a large language model (LLM) and prompt engineering. Front Commun.8:1. (https://doi.org/10.3389/fcomm.2023.1129082).

Lin YL et al. (2025). Knowledge-point classification using simple LSTM-based and siamese-based networks for virtual patient simulation. BMC Med Inform Decis Mak.25(1):39. (https://doi.org/10.1186/s12911-025-02866-3).

Ling Kuo RY et al. (2024). Stakeholder perspectives towards diagnostic artificial intelligence: a co-produced qualitative evidence synthesis. EClinicalMedicine.71:102555. (https://doi.org/10.1016/j.eclinm.2024.102555).

Litvinova Y et al. (2025). Personalized medicine for healthier populations: key considerations for policy-makers. Copenhagen: European Observatory on Health Systems and Policies (https://eurohealthobservatory.who.int/publications/i/personalized-medicine-for-healthier-populations-key-considerations-for-policy-makers).

Liu F et al. (2023a). A medical multimodal large language model for future pandemics. NPJ Digit Med.6(1):226. (https://doi.org/10.1038/s41746-023-00952-2).

Liu M et al. (2024a). FAIM: fairness-aware interpretable modeling for trustworthy machine learning in healthcare. Patterns (N Y).5(10):101059. (https://doi.org/10.1016/j.patter.2024.101059).

Liu P et al. (2023b). A scoping review of the clinical application of machine learning in data-driven population segmentation analysis. J Am Med Inform Assoc.30(9):1573–82. (https://doi.org/10.1093/jamia/ocad111).

Liu X et al. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med.26(9):1364–74. (https://doi.org/10.1038/s41591-020-1034-x).

Liu X et al. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health.1(6):e271–e97. (https://doi.org/10.1016/S2589-7500(19)30123-2).

Liu Y et al. (2024b). Evolution of surgical robot systems enhanced by artificial intelligence: a review. Advanced Intelligent Systems.6(5). (https://doi.org/10.1002/aisy.202300268).

Lo LS (2023). The CLEAR path: a framework for enhancing information literacy through prompt engineering. J Acad Librarianship.49(4):102720. (https://doi.org/10.1016/j.acalib.2023.102720).

Lopez de Coca T et al. (2022). Bridging the generational digital divide in the healthcare environment. J Pers Med.12(8):1214. (https://doi.org/10.3390/jpm12081214).

Lu J et al. (2024a). An environmental uncertainty perception framework for misinformation detection and spread prediction in the COVID-19 pandemic: artificial intelligence approach. JMIR AI.3:e47240. (https://doi.org/10.2196/47240).

Lu MT et al. (2020). Deep learning using chest radiographs to identify high-risk smokers for lung cancer screening computed tomography: development and validation of a prediction model. Ann Intern Med.173(9):704–13. (https://doi.org/10.7326/M20-1868).

Lu MY et al. (2024b). A visual-language foundation model for computational pathology. Nat Med.30(3):863–74. (https://doi.org/10.1038/s41591-024-02856-4).

Lu W et al. (2025). Fine-tuning large language models for domain adaptation: exploration of training strategies, scaling, model merging and synergistic capabilities. npj Comp Mater.11(1). (https://doi.org/10.1038/s41524-025-01564-y).

Lukyanenko R et al. (2022). Trust in artificial intelligence: from a Foundational Trust Framework to emerging research opportunities. Electronic Markets.32(4):1993–2020. (https://doi.org/10.1007/s12525-022-00605-4).

Lunežnik P (2024). Pregled praks sprejemanja klicev v ambulantah družinske medicine Ministrstvo za zdravje Republike Slovenije.

Ma R et al. (2020). Machine learning in the optimization of robotics in the operative field. Curr Opin Urol.30(6):808–16. (https://doi.org/10.1097/MOU.0000000000000816).

Maleki Varnosfaderani S, Forouzanfar M (2024). The role of AI in hospitals and clinics: transforming healthcare in the 21st century. Bioengineering (Basel).11(4):337. (https://doi.org/10.3390/bioengineering11040337).

Marcus HJ et al. (2024). The IDEAL framework for surgical robotics: development, comparative evaluation and long-term monitoring. Nat Med.30(1):61–75. (https://doi.org/10.1038/s41591-023-02732-7).

Marko JGO et al. (2025). Examining inclusivity: the use of AI and diverse populations in health and social care: a systematic review. BMC Med Inform Decis Mak.25(1):57. (https://doi.org/10.1186/s12911-025-02884-1).

Marteau TM et al. (2021). Changing behaviour: an essential component of tackling health inequalities. BMJ.372:n332. (https://doi.org/10.1136/bmj.n332).

Martí T, González López-Valcárcel B (2025). Hacia la transformación del sistema de salud catalán: CAIROS y la atención primaria. Nada es Gratis. 10 April (https://nadaesgratis.es/beatriz-gonzalez-lopez-valcarcel/hacia-la-transformacion-del-sistema-de-salud-catalan-cairos-y-la-atencion-primaria).

Maslej N et al. (2025). Artificial intelligence index report 2025. Stanford, CA: Institute for Human-Centered AI, Stanford University (https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf).

Matthay EC et al. (2025). Integrating artificial intelligence into causal research in epidemiology. Current Epidemiology Reports.12(1):6. (https://doi.org/10.1007/s40471-025-00359-5).

McCarthy J et al. (1955). A proposal for the Dartmouth Summer Research Project on Artificial Intelligence. AI Magazine.27(4):12.

McKee M, Correia T (2025). The future of the health professions: navigating shortages, imbalances, and automation. Int J Health Plann Manage.40(2):289–92. (https://doi.org/10.1002/hpm.3865).

McKee M et al. (2025a). A public health response to economic warfare. Int J Health Plann Manage.40(5):1025–8. (https://doi.org/10.1002/hpm.3940).

McKee M et al. (2025b). The power of artificial intelligence for managing pandemics: a primer for public health professionals. Int J Health Plann Manage.40(1):257–70. (https://doi.org/10.1002/hpm.3864).

McKee M et al. (2024). Trust: the foundation of health systems. Copenhagen: WHO Regional Office for Europe.

McKinney SM et al. (2020). International evaluation of an AI system for breast cancer screening. Nature.577(7788):89–94. (https://doi.org/10.1038/s41586-019-1799-6).

McLean S et al. (2013). The impact of telehealthcare on the quality and safety of care: a systematic overview. PLoS One.8(8):e71238. (https://doi.org/10.1371/journal.pone.0071238).

Meacham S (2023). A race to extinction: how great power competition is making artificial intelligence existentially dangerous. Harvard International Review. 8 Sept (https://hir.harvard.edu/a-race-to-extinction-how-great-power-competition-is-making-artificial-intelligence-existentially-dangerous/).

Mehari M et al. (2025). Predicting therapeutic clinical trial enrollment for adult patients with low- and high-grade glioma using supervised machine learning. Sci Adv.11(23):eadt5708. (https://doi.org/10.1126/sciadv.adt5708).

Mehl G et al. (2021). WHO SMART guidelines: optimising country-level use of guideline recommendations in the digital age. Lancet Digit Health.3(4):e213–e6. (https://doi.org/10.1016/S2589-7500(21)00038-8).

Mengüç K et al. (2023). Optimisation of COVID-19 vaccination process using GIS, machine learning, and the multi-layered transportation model. Int J Prod Res.63(2):404–17. (https://doi.org/10.1080/00207543.2023.2182151).

Microsoft (2025). Seeing AI [website]. Microsoft. (https://www.microsoft.com/en-us/garage/wall-of-fame/seeing-ai/).

Miller T (2019). Explanation in artificial intelligence: insights from the social sciences. Artificial Intelligence.267:1–38. (https://doi.org/10.1016/j.artint.2018.07.007).

Mittermaier M et al. (2023). Bias in AI-based models for medical applications: challenges and mitigation strategies. NPJ Digit Med.6(1):113. (https://doi.org/10.1038/s41746-023-00858-z).

Mokander J et al. (2022). Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European ai regulation. Minds Mach (Dordr).32(2):241–68. (https://doi.org/10.1007/s11023-021-09577-4).

Moreno-Torres JG et al. (2012). A unifying view on dataset shift in classification. Pattern Recognit.45(1):521–30. (https://doi.org/10.1016/j.patcog.2011.06.019).

Moret M et al. (2023). Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. Nat Commun.14(1):114. (https://doi.org/10.1038/s41467-022-35692-6).

Morgan M, Chinn S (1983). ACORN group, social class, and child health. J Epidemiol Community Health.37(3):196–203. (https://doi.org/10.1136/jech.37.3.196).

Morley J et al. (2025). Can a digital NHS be equitable? BMJ.389:r1317. (https://doi.org/10.1136/bmj.r1317).

Morley J et al. (2020). The ethics of AI in health care: a mapping review. Soc Sci Med.260:113172. (https://doi.org/10.1016/j.socscimed.2020.113172).

Mosqueira-Rey E et al. (2022). Human-in-the-loop machine learning: a state of the art. Artif Intell Rev.56(4):3005–54. (https://doi.org/10.1007/s10462-022-10246-w).

Moulds A, Horton T (2023). What do technology and AI mean for the future of work in health care?, London: The Health Foundation (https://www.health.org.uk/reports-and-analysis/briefings/what-do-technology-and-ai-mean-for-the-future-of-work-in-health-care).

Muldoon J, Wu BA (2023). Artificial intelligence in the colonial matrix of power. Philos Technol.36(4). (https://doi.org/10.1007/s13347-023-00687-8).

Mulligan DK, Bamberger KA (2018). Saving governance-by-design. California Law Review.106(3):697–784.

Mulrow CD (1994). Rationale for systematic reviews. BMJ.309(6954):597–9. (https://doi.org/10.1136/bmj.309.6954.597).

Murdoch B (2021). Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Med Ethics.22(1):122. (https://doi.org/10.1186/s12910-021-00687-3).

Narajala VS, Narayan O (2025). Securing agentic AI: a comprehensive threat model and mitigation framework for generative ai agents. arXiv.2504:19956. (https://doi.org/10.48550/arXiv.2504.19956).

Narayanan A (2023). Understanding social media recommendation algorithms. Knight First Amendment Institute. 9 March (https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms).

National Academies of Science Engineering and Medicine (2025). The age of AI in the life sciences: benefits and biosecurity considerations. Washington, DC: National Academies Press (https://doi.org/10.17226/28868).

National Academy of Medicine (2025). Generative artificial intelligence in health and medicine: opportunities and responsibilities for transformative innovation. Washington, DC: The National Academies Press (https://doi.org/10.17226/28907).

National Institute of Standards and Technology (2024). Artificial intelligence risk management framework: generative artificial intelligence profile [website]. National Institute of Standards and Technology. (https://www.nist.gov/itl/ai-risk-management-framework).

Nature (2025). Don't sleepwalk from computer-vision research into surveillance. Nature.642(8069):839–40. (https://doi.org/10.1038/d41586-025-01965-5).

Nature Electronics (2025). Editorial: a new model for AI. Nat Electron.8(2):95. (https://doi.org/10.1038/s41928-025-01361-x).

Naughton F et al. (2023). An automated, online feasibility randomized controlled trial of a just-in-time adaptive intervention for smoking cessation (Quit Sense). Nicotine Tob Res.25(7):1319–29. (https://doi.org/10.1093/ntr/ntad032).

Nguyen H et al. (2024). A comparative study of quality evaluation methods for text summarization. arXiv.2407.00747 (https://doi.org/10.48550/arXiv.2407.00747).

NHS Digital (2025). London Secure Data Environment [website]. NHS England. (https://digital.nhs.uk/data-and-information/research-powered-by-data/sde-network/london-secure-data-environment).

NHS England (2018). Population health management flatpack. London: NHS England (https://imperialcollegehealthpartners.com/wp-content/uploads/2018/07/Population-Health-Management-Flatpack-Version-1.0-Final-Sent.pdf).

NHS England (2021). Digitial inclusion in healthcare [website]. NHS England. (https://www.england.nhs.uk/ltphimenu/digital-inclusion/digital-inclusion-in-health-and-care/).

NHS England (2025). AI-enabled ambient scribing products in health and care settings [website]. NHS England. (https://www.england.nhs.uk/long-read/ai-enabled-ambient-scribing-products-in-health-and-care-settings/).

NHS Transformation Directorate (2022). Using an AI chatbot to streamline mental health referrals [website]. NHS England. (https://transform.england.nhs.uk/key-tools-and-info/digital-playbooks/workforce-digital-playbook/using-an-ai-chatbot-to-streamline-mental-health-referrals/).

NICE (2022). Digital technologies for the detection of melanoma. London: National Institute for Health and Care Excellence (https://www.nice.org.uk/advice/mib311).

NICE (2025). Use of AI in evidence generation: NICE position statement. London: National Institute for Health and Care Excellence (https://www.nice.org.uk/about/what-we-do/our-research-work/use-of-ai-in-evidence-generation--nice-position-statement).

Noori K (2024). A disability-inclusive Artificial Intelligence Act: a guide to monitor implementation in your country. Brussels: European Disability Forum (https://www.edf-feph.org/publications/a-disability-inclusive-artificial-intelligence-act-a-guide-to-monitor-implementation-in-your-country/).

Norwood AA (2025). Understanding AI-enabled biological threats: hype, hazard, and governance. Engineering Biology in Cambridge. 15 April (https://www.engbio.cam.ac.uk/news/understanding-ai-enabled-biological-threats-hype-hazard-and-governance).

Novelli C et al. (2023). Accountability in artificial intelligence: what it is and how it works. AI Soc.39(4):1871–82. (https://doi.org/10.1007/s00146-023-01635-y).

Nuclear Threat Initiative (2025). Statement on biosecurity risks at the convergence of AI and the life sciences. Nuclear Threat Initiative. 17 July (https://www.nti.org/analysis/articles/statement-on-biosecurity-risks-at-the-convergence-of-ai-and-the-life-sciences/).

O'Donnell J (2025). How a new type of AI is helping police skirt facial recognition bans. Artificial Intelligence. 12 May (https://www.technologyreview.com/2025/05/12/1116295/how-a-new-type-of-ai-is-helping-police-skirt-facial-recognition-bans/).

O'Grady E (2024). Science & Tech: Why AI fairness conversations must include disabled people. The Harvard Gazette. 3 April (https://news.harvard.edu/gazette/story/2024/04/why-ai-fairness-conversations-must-include-disabled-people/).

Oddy C et al. (2024). Promising algorithms to perilous applications: a systematic review of risk stratification tools for predicting healthcare utilisation. BMJ Health Care Inform.31(1):e101065. (https://doi.org/10.1136/bmjhci-2024-101065).

OECD (2020). Tracking and tracing COVID: Protecting privacy and data while using apps and biometrics. OECD Policy Responses to Coronavirus (COVID-19). Paris: OECD Publishing (https://doi.org/10.1787/8f394636-en.).

OECD.AI Policy Observatory (2025). OECD AI principles overview [website]. OECD. (https://oecd.ai/en/ai-principles).

Okidegbe N (2022). The democratizing potential of algorithms?, 53 Connecticut Law Review. 739: (https://scholarship.law.bu.edu/faculty_scholarship/3138).

Olawade DB et al. (2023). Using artificial intelligence to improve public health: a narrative review. Front Public Health.11:1196397. (https://doi.org/10.3389/fpubh.2023.1196397).

One London (2025). A framework for the safe, efficient and effective implementation, use and maintenance of AI in health and care in London., London: NHS (https://www.onelondon.online/wp-content/uploads/2025/03/A-Framework-for-the-safe-efficient-and-effective-implementation-use-and-maintenance-of-AI-in-health-and-care-in-London.pdf).

ONS (2019). Exploring the UK's digital divide. London: Office for National Statistics (https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/articles/exploringtheuksdigitaldivide/2019-03-04#what-is-the-pattern-of-digital-exclusion-across-the-uk).

OpenAI (2025a). How ChatGPT and our foundation models are developed [website]. OpenAI. (https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed).

OpenAI (2025b). Introducing HealthBench [website]. OpenAI. (https://openai.com/index/healthbench/).

OpenSAFELY (2025). Secure analytics platform for NHS electronic health records [website]. OpenSAFELY. (https://www.opensafely.org/).

Orhan F, Kurutkan MN (2025). Predicting total healthcare demand using machine learning: separate and combined analysis of predisposing, enabling, and need factors. BMC Health Serv Res.25(1):366. (https://doi.org/10.1186/s12913-025-12502-5).

Oura (2025). Introducing Oura Advisor: your AI-powered personal health companion [website]. Oura. (https://ouraring.com/blog/oura-advisor/).

Panagopoulos G et al. (2021). Transfer graph neural networks for pandemic forecasting. Proceedings of the AAAI Conference on Artificial Intelligence.35(6):4838–45. (https://doi.org/10.1609/aaai.v35i6.16616).

Pancholi S et al. (2024). Use of artificial intelligence techniques to assist individuals with physical disabilities. Annu Rev Biomed Eng.26(1):1–24. (https://doi.org/10.1146/annurev-bioeng-082222-012531).

Panteli D et al. (2025). Artificial intelligence in public health: promises, challenges, and an agenda for policy makers and public health institutions. Lancet Public Health.10(5):e428–e32. (https://doi.org/10.1016/S2468-2667(25)00036-2).

Panteli D et al. (2024). Artificial intelligence in public health: lessons from the EPH Conference. Eurohealth.30(3):13–8.

Pastika L et al. (2024). Abstract 4137169: artificial intelligence-enabled electrocardiography for the prediction of future type 2 diabetes mellitus. Circulation.150(Suppl_1):A4137169–A41371. (https://doi.org/10.1161/circ.150.suppl_1.4137169).

Patra BG et al. (2021). Extracting social determinants of health from electronic health records using natural language processing: a systematic review. J Am Med Inform Assoc.28(12):2716–27. (https://doi.org/10.1093/jamia/ocab170).

Pelekis S et al. (2025). Adversarial machine learning: a review of methods, tools, and critical industry sectors. Artif Intell Rev.58(8):226. (https://doi.org/10.1007/s10462-025-11147-4).

Permanent Representation of Belgium to the European Union (2024). Citizen's panel on AI issues. 22 May (https://europeanunion.diplomatie.belgium.be/en/citizens-panel-ai-issues).

Perski O et al. (2024). Supervised machine learning to predict smoking lapses from Ecological Momentary Assessments and sensor data: implications for just-in-time adaptive intervention development. PLOS Digit Health.3(8):e0000594. (https://doi.org/10.1371/journal.pdig.0000594).

Petersen M et al. (2024). Artificial intelligence-based copilots to generate causal evidence. New Eng J Med AI.1(12). (https://doi.org/10.1056/AIp2400727).

Pianykh OS et al. (2020). Improving healthcare operations management with machine learning. Nat Mach Intell.2(5):266–73. (https://doi.org/10.1038/s42256-020-0176-3).

Pilati F, Venturini T (2025). The use of artificial intelligence in counter-disinformation: a world wide (web) mapping. Front Polit Sci.7:1517726. (https://doi.org/10.3389/fpos.2025.1517726).

Ping L et al. (2023). Application and evaluation of surgical tool and tool tip recognition based on Convolutional Neural Network in multiple endoscopic surgical scenarios. Surg Endosc.37(9):7376–84. (https://doi.org/10.1007/s00464-023-10323-3).

Podder S (2024). Reflections on the first binding regulation on AI globally. San Francisco: Green Software Foundation (https://greensoftware.foundation/articles/the-eu-ai-act-insights-from-the-green-ai-committee).

Pol.is (2025). Input crowd, output meaning [website]. Polis. (https://pol.is/home).

POST (2024). Artificial intelligence (AI) glossary [website]. Parliamentary Office of Science and Technology, United Kingdom Parliament. (https://post.parliament.uk/artificial-intelligence-ai-glossary/).

Pöysti T (2018). Trust on digital administration and platforms. Scand Stud Law.65:321–63.

Pöysti T (2023). Legislating for legal certainty, with a Right to a Human Face, in an automated public administration. In: Suksi M, editor. The rule of law and automated decision-making: exploring fundamentals of algorithmic governance. Cham: Springer Nature: 33–63.

Pöysti T (2024). The precautionary approach design pattern. Digital Society.3(1):5. (https://doi.org/10.1007/s44206-024-00090-6).

Preiksaitis C et al. (2024). The role of large language models in transforming emergency medicine: scoping review. JMIR Med Inform.12:e53787. (https://doi.org/10.2196/53787).

Preiksaitis C, Rose C (2023). Opportunities, challenges, and future directions of generative artificial intelligence in medical education: Scoping review. JMIR Med Educ.9:e48785. (https://doi.org/10.2196/48785).

Price WN, 2nd et al. (2019). Potential liability for physicians using artificial intelligence. JAMA.322(18):1765–6. (https://doi.org/10.1001/jama.2019.15064).

pwc (2025). The fearless future: 2025 global AI jobs barometer [website]. pwc. (https://www.pwc.com/gx/en/issues/artificial-intelligence/ai-jobs-barometer.html).

Qian C, Ren H (2025). Deep reinforcement learning in surgical robotics: enhancing the automation level. In: Zequi SdC, Hongliang R, editors. Handbook of robotic surgery. London: Academic Press: 89–102.

Qin L et al. (2025). A survey of multilingual large language models. Patterns (N Y).6(1):101118. (https://doi.org/10.1016/j.patter.2024.101118).

Radanliev P (2025). AI ethics: integrating transparency, fairness, and privacy in AI development. Appl Artif Intell.39(1):2463722. (https://doi.org/10.1080/08839514.2025.2463722).

Rajkomar A et al. (2019). Machine learning in medicine. New Eng J Med.380(14):1347–58. (https://doi.org/10.1056/NEJMra1814259).

Ramezani M et al. (2025). Applications of artificial intelligence and the challenges in health technology assessment: a scoping review and framework with a focus on economic dimensions. Health Econ Rev.15(1):46. (https://doi.org/10.1186/s13561-025-00645-4).

Reicher S (2024). Trust the public? How the UK Government got the psychology of COVID-19 wrong and why it matters. In: Pasternak C, editor. Evaluating a pandemic. Singapore: World Scientific: 98–118.

Republic of Slovenia (2021). National Programme to Promote the Development and Use of Artificial Intelligence in the Republic of Slovenia by 2025 (NpAI) [website].(https://url.uk.m.mimecastprotect.com/s/btHQCoy14umNvNmI1frhp_wsQ?domain=gov.si).

Republic of Slovenia (2023). Digitalna Slovenija 2030 [website].(https://www.gov.si/assets/ministrstva/MDP/Dokumenti/DSI2030-potrjena-na-Vladi-RS_marec-2023.pdf).

Resch B et al. (2025). The generative revolution: AI foundation models in geospatial health-applications, challenges and future research. Int J Health Geogr.24(1):6. (https://doi.org/10.1186/s12942-025-00391-0).

Reverberi C et al. (2022). Experimental evidence of effective human-AI collaboration in medical decision-making. Sci Rep.12(1):14952. (https://doi.org/10.1038/s41598-022-18751-2).

Rieke N et al. (2020). The future of digital health with federated learning. NPJ Digit Med.3(1):119. (https://doi.org/10.1038/s41746-020-00323-1).

Riley BK, Dixon A (2024). Emotional and cognitive trust in artificial intelligence: a framework for identifying research opportunities. Curr Opin Psychol.58:101833. (https://doi.org/10.1016/j.copsyc.2024.101833).

Rimal RN, Lapinski MK (2009). Why health communication is important in public health. Bull World Health Organ.87(4):247. (https://doi.org/10.2471/blt.08.056713).

Rodilosso E (2024). Filter bubbles and the unfeeling: how AI for social media can foster extremism and polarization. Philos Technol.37(2):71. (https://doi.org/10.1007/s13347-024-00758-4).

Rodriguez-Ruiz A et al. (2019). Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology.290(2):305–14. (https://doi.org/10.1148/radiol.2018181371).

Rogers EM (1962). Diffusion of innovations. New York: The Free Press.

Rosenbacke R et al. (2024a). How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: systematic review. JMIR AI.3:e53207. (https://doi.org/10.2196/53207).

Rosenbacke R et al. (2024b). False conflict and false confirmation errors are crucial components of AI accuracy in medical decision making. Nat Commun.15(1):6896. (https://doi.org/10.1038/s41467-024-50952-3).

Ru X et al. (2024). Identify potential drug candidates within a high-quality compound search space. Brief Bioinform.26(1):bbaf024. (https://doi.org/10.1093/bib/bbaf024).

Ruberto AJ et al. (2021). The future of simulation-based medical education: adaptive simulation utilizing a deep multitask neural network. AEM Educ Train.5(3):e10605. (https://doi.org/10.1002/aet2.10605).

Saeidi H et al. (2022). Autonomous robotic laparoscopic surgery for intestinal anastomosis. Sci Robot.7(62):eabj2908. (https://doi.org/10.1126/scirobotics.abj2908).

Safranek CW et al. (2023). The role of large language models in medical education: applications and implications. JMIR Med Educ.9:e50945. (https://doi.org/10.2196/50945).

Sahni N et al. (2023). The potential impact of artificial intelligence on healthcare spending. Cambridge, MA: National Bureau of Economic Research (https://dx.doi.org/10.3386/w30857).

Saito T, Rehmsmeier M (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One.10:e0118432. (https://doi.org/10.1371/journal.pone.0118432).

Salinas MP et al. (2024). A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis. NPJ Digit Med.7(1):125. (https://doi.org/10.1038/s41746-024-01103-x).

Samborska V (2025). Scaling up: how increasing inputs has made artificial intelligence more capable. Our World in Data. 20 Jan (https://ourworldindata.org/scaling-up-ai).

Sanders NE, Schneier B (2024). Let's not make the same mistakes with AI that we made with social media. Business. 13 March (https://www.technologyreview.com/2024/03/13/1089729/lets-not-make-the-same-mistakes-with-ai-that-we-made-with-social-media/).

Sarker A et al. (2024). Natural language processing for digital health in the era of large language models. Yearb Med Inform.33(1):229–40. (https://doi.org/10.1055/s-0044-1800750).

Schmalzle R, Wilcox S (2022). Harnessing artificial intelligence for health message generation: the folic acid message engine. J Med Internet Res.24(1):e28858. (https://doi.org/10.2196/28858).

Schmid S et al. (2025). Arms race or innovation race? Geopolitical AI development. Geopolitics.30(4):1907–36. (https://doi.org/10.1080/14650045.2025.2456019).

Schramowski P et al. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. Nat Mach Intell.4(3):258–68. (https://doi.org/10.1038/s42256-022-00458-8).

Schulte-Sasse R et al. (2021). Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. Nat Mach Intell.3(6):513–26. (https://doi.org/10.1038/s42256-021-00325-y).

Sekalala S et al. (2020). Health and human rights are inextricably linked in the COVID-19 response. BMJ Glob Health.5(9):e003359. (https://doi.org/10.1136/bmjgh-2020-003359).

Sendak MP et al. (2020). Real-World Integration of a Sepsis Deep Learning Technology Into Routine Clinical Care: Implementation Study. JMIR Med Inform.8(7):e15182. (https://doi.org/10.2196/15182).

Shahid S et al. (2025). Diagnostic accuracy of apple watch electrocardiogram for atrial fibrillation: a systematic review and meta-analysis. JACC Adv.4(2):101538. (https://doi.org/10.1016/j.jacadv.2024.101538).

Shashkevich A (2019). Stanford researcher examines earliest concepts of artificial intelligence, robots in ancient myths. Arts & Humanities. 28 Feb (https://news.stanford.edu/stories/2019/02/ancient-myths-reveal-early-fantasies-artificial-life).

Shayegh S et al. (2023). Prioritizing COVID-19 vaccine allocation in resource poor settings: towards an artificial intelligence-enabled and geospatial-assisted decision support framework. PLoS One.18(8):e0275037. (https://doi.org/10.1371/journal.pone.0275037).

Sheng B et al. (2024). Artificial intelligence for diabetes care: current and future prospects. Lancet Diabetes Endocrinol.12(8):569–95. (https://doi.org/10.1016/S2213-8587(24)00154-2).

Shortliffe EH, Sepulveda MJ (2018). Clinical Decision Support in the Era of Artificial Intelligence. JAMA.320(21):2199–200. (https://doi.org/10.1001/jama.2018.17163).

Shumailov I et al. (2024). AI models collapse when trained on recursively generated data. Nature.631(8022):755–9. (https://doi.org/10.1038/s41586-024-07566-y).

Siabi N et al. (2024). Comprehensive investigation of climatic and socio-economic factors on dengue fever prediction using machine learning approaches. In: Proceedings. AGU24: Whats's next for science, Washington, DC, 9–13 December. Washington, DC: American Geophysical Union; GH52B–02 (https://agu.confex.com/agu/agu24/meetingapp.cgi/Paper/1589706).

Siddiqui TAA et al. (2024). Climate-driven future habitat prediction of arbovirus vectors Aedes aegypti and Aedes albopictus. In: Proceedings. AGU24: Whats's next for science, Washington, DC, 9–13 December. Washington, DC: American Geophysical Union; GH52B–01 (https://ui.adsabs.harvard.edu/abs/2024AGUFMGH52B..01S/abstract).

Siemens W et al. (2025). Opportunities, challenges and risks of using artificial intelligence for evidence synthesis. BMJ Evid Based Med.bmjebm–2024–113. (https://doi.org/10.1136/bmjebm-2024-113320).

Singh N et al. (2023). Drug discovery and development: introduction to the general public and patient groups. Front Drug Discov.3:461. (https://doi.org/10.3389/fddsv.2023.1201419).

Singhal K et al. (2025). Toward expert-level medical question answering with large language models. Nat Med.31(3):943–50. (https://doi.org/10.1038/s41591-024-03423-7).

Siontis GCM et al. (2021). Development and validation pathways of artificial intelligence tools evaluated in randomised clinical trials. BMJ Health Care Inform.28(1). (https://doi.org/10.1136/bmjhci-2021-100466).

SmartCHANGE (2025). Empowering youth: AI models for a healthier future [website]. SmartCHANGE. (https://smart-change.eu/).

Sohn R, Stix G (2022). AI drug discovery systems might be repurposed to make chemical weapons, researchers warn. Scientific American. 21 April (https://www.scientificamerican.com/article/ai-drug-discovery-systems-might-be-repurposed-to-make-chemical-weapons-researchers-warn/).

Song Y et al. (2023). Advances in geocomputation and geospatial artificial intelligence (GeoAI) for mapping. Int J Appl Earth Obs Geoinf.120:103300. (https://doi.org/10.1016/j.jag.2023.103300).

Sounderajah V et al. (2021). A national survey assessing public readiness for digital health strategies against COVID-19 within the United Kingdom. Sci Rep.11(1):5958. (https://doi.org/10.1038/s41598-021-85514-w).

Stamer T et al. (2023). Artificial intelligence supporting the training of communication skills in the education of health care professions: scoping review. J Med Internet Res.25:e43311. (https://doi.org/10.2196/43311).

Stanford University (2024). BioMedLM [website]. Stanford University. (https://crfm.stanford.edu/2022/12/15/biomedlm.html).

Stevens N, Keyes O (2021). Seeing infrastructure: race, facial recognition and the politics of data. Cult Stud.35(4-5):833–53. (https://doi.org/10.1080/09502386.2021.1895252).

Stuke H et al. (2025). Peer relationships are a direct cause of the adolescent mental health crisis: interpretable machine learning analysis of 2 large cohort studies. JMIR Public Health Surveill.11:e60125. (https://doi.org/10.2196/60125).

Sun Y et al. (2024). AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. Humanit Soc Sci.11(1):1278. (https://doi.org/10.1057/s41599-024-03811-x).

Sung J, Hopper JL (2023). Co-evolution of epidemiology and artificial intelligence: challenges and opportunities. Int J Epidemiol.52(4):969–73. (https://doi.org/10.1093/ije/dyad089).

Swartout WR (1985). Rule-based expert systems: the mycin experiments of the Stanford heuristic programming project. Artificial Intelligence.26(3):364–6. (https://doi.org/10.1016/0004-3702(85)90067-0).

Taherdoost H (2023). Deep learning and neural networks: decision-making implications. Symmetry.15(9):1723. (https://doi.org/10.3390/sym15091723).

Tam TYC et al. (2024). A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med.7(1):258. (https://doi.org/10.1038/s41746-024-01258-7).

Tarumi S et al. (2021). Leveraging artificial intelligence to improve chronic disease care: methods and application to pharmacotherapy decision support for type-2 diabetes mellitus. Methods Inf Med.60(S 01):e32–e43. (https://doi.org/10.1055/s-0041-1728757).

Thadani NN et al. (2023). Learning from prepandemic data to forecast viral escape. Nature.622(7984):818–25. (https://doi.org/10.1038/s41586-023-06617-0).

The Framingham Heart Study (2025). Cardiovascular Disease (10-year risk) [website]. The Framingham Heart Study. (https://www.framinghamheartstudy.org/fhs-risk-functions/cardiovascular-disease-10-year-risk/).

Thieme A et al. (2023). Designing human-centered AI for mental health: developing clinically relevant applications for online CBT treatment. ACM Trans Comput-Hum Interact.30(2):1–50. (https://doi.org/10.1145/3564752).

Tierney AA et al. (2025). Ambient artificial intelligence scribes: learnings after 1 year and over 2.5 million uses. NEJM Catalyst.6:5. (https://doi.org/10.1056/cat.25.0040).

Tomasev N et al. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. Nature.572(7767):116–9. (https://doi.org/10.1038/s41586-019-1390-1).

Topol EJ (2019). High-performance medicine: the convergence of human and artificial intelligence. Nat Med.25(1):44–56. (https://doi.org/10.1038/s41591-018-0300-7).

Tornatzky L, Fleischer M (1990). The process of technology innovation. Lexington, MA: Lexington Books.

Truhn D et al. (2024). Large language models and multimodal foundation models for precision oncology. NPJ Precis Oncol.8(1):72. (https://doi.org/10.1038/s41698-024-00573-2).

Tseng AS et al. (2021). Spectrum bias in algorithms derived by artificial intelligence: a case study in detecting aortic stenosis using electrocardiograms. Eur Heart J Digit Health.2(4):561–7. (https://doi.org/10.1093/ehjdh/ztab061).

Tudor Car L et al. (2020). Conversational agents in health care: scoping review and conceptual analysis. J Med Internet Res.22(8):e17158. (https://doi.org/10.2196/17158).

Turing AM (1950). I.—Computing machinery and intelligence. Mind.LIX(236):433–60. (https://doi.org/10.1093/mind/LIX.236.433).

UCL (2025). AI model trained on de-identified data from 57 million people. UCL news. 7 May (https://www.ucl.ac.uk/news/2025/may/ai-model-trained-de-identified-data-57-million-people).

Uebler U (2001). Multilingual speech recognition in seven languages. Speech Commun.35(1-2):53–69. (https://doi.org/10.1016/s0167-6393(00)00095-9).

United Kingdom Government (2025). Fit for the future: the 10 year health plan for England. London: The Stationery Office (https://assets.publishing.service.gov.uk/media/6888a0b1a11f859994409147/fit-for-the-future-10-year-health-plan-for-england.pdf).

UK Health Security Agency (2025). How we are pioneering artificial intelligence applications in public health. London: UK Health Security Agency (https://ukhsa.blog.gov.uk/2025/03/14/how-we-are-pioneering-artificial-intelligence-applications-in-public-health/).

UK Research Integrity Office (2025). Embracing AI with integrity. A practical guide for researchers. London: UK Research Integrity Office (https://ukrio.org/ukrio-resources/embracing-ai-with-integrity/).

Umbrello S, van de Poel I (2021). Mapping value sensitive design onto AI for social good principles. AI Ethics.1(3):283–96. (https://doi.org/10.1007/s43681-021-00038-3).

UNEP (2024). AI has an environmental problem. Here what the world can do about that. Environment under review. 21 Sept (https://www.unep.org/news-and-stories/story/ai-has-environmental-problem-heres-what-world-can-do-about).

UNESCO (2021). Recommendation on the ethics of artificial intelligence. Paris: United Nations Educational, Scientific Cultural Organization (https://unesdoc.unesco.org/notice?id=p::usmarcdef_0000381137).

USAID Center for Innovation and Impact et al. (2019). Artificial intelligence in global health: defining a collective path forward. Washington, DC: USAID (https://www.usaid.gov/cii/ai-global-health).

Valarmathi B et al. (2024). Sentiment analysis of Covid-19 Twitter data using deep learning algorithm. Procedia Comput Sci.235:3397–407. (https://doi.org/10.1016/j.procs.2024.04.320).

van Buchem MM et al. (2021). The digital scribe in clinical practice: a scoping review and research agenda. NPJ Digit Med.4(1):57. (https://doi.org/10.1038/s41746-021-00432-5).

van Deursen AJ (2020). Digital inequality during a pandemic: quantitative study of differences in COVID-19–related internet uses and outcomes among the general population. J Med Internet Res.22(8):e20073. (https://doi.org/10.2196/20073).

Vasey B et al. (2023). Intraoperative applications of artificial intelligence in robotic surgery: a scoping review of current development stages and levels of autonomy. Ann Surg.278(6):896–903. (https://doi.org/10.1097/SLA.0000000000005700).

Vaswani A et al. (2017). Attention is all you need. Long Beach, CA: Curran Associates Inc.

Volinsky-Fremond S et al. (2024). Prediction of recurrence risk in endometrial cancer with multimodal deep learning. Nat Med.30(7):1962–73. (https://doi.org/10.1038/s41591-024-02993-w).

Waldman A (2024). How United Health's playbook for limiting mental health coverage puts countless Americans' treatment at risk. Health Care. 19 Nov (https://www.propublica.org/article/unitedhealth-mental-health-care-denied-illegal-algorithm).

Wang F et al. (2023a). Surgical smoke removal via residual Swin transformer network. Int J Comput Assist Radiol Surg.18(8):1417–27. (https://doi.org/10.1007/s11548-023-02835-z).

Wang G et al. (2023b). Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. Nat Med.29(10):2633–42. (https://doi.org/10.1038/s41591-023-02552-9).

Wang J et al. (2024). High-resolution modeling and projection of heat-related mortality in Germany under climate change. Commun Med (Lond).4(1):206. (https://doi.org/10.1038/s43856-024-00643-3).

Wang L et al. (2021). Artificial intelligence for COVID-19: a systematic review. Front Med (Lausanne).8:704256. (https://doi.org/10.3389/fmed.2021.704256).

Wang Y et al. (2022). Understanding and neutralising covid-19 misinformation and disinformation. BMJ.379:e070331. (https://doi.org/10.1136/bmj-2022-070331).

Watson JL et al. (2023). De novo design of protein structure and function with RFdiffusion. Nature.620(7976):1089–100. (https://doi.org/10.1038/s41586-023-06415-8).

Wehrli S et al. (2024). The role of the (in)accessibility of social media data for infodemic management: a public health perspective on the situation in the European Union in March 2024. Front Public Health.12:1378412. (https://doi.org/10.3389/fpubh.2024.1378412).

Weiser B, Schweber N (2023). The ChatGPT lawyer explains himself. The New York Times. 8 June (https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html).

WHO (2021). Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization (https://iris.who.int/handle/10665/34199).

WHO (2024). Questions and answers: determinants of health [website]. World Health Organization. (https://www.who.int/news-room/questions-and-answers/item/determinants-of-health).

WHO (2025a). Behavioural sciences for better health [website]. World Health Organization. (https://www.who.int/initiatives/behavioural-sciences).

WHO (2025b). Health technology assessment [website]. World Health Organization. (https://www.who.int/health-topics/health-technology-assessment).

WHO (2025c). S.A.R.A.H, a Smart AI Resource Assistant for Health [website]. World Health Organization. (https://www.who.int/campaigns/s-a-r-a-h).

WHO (2025d). Social determinants of health [website]. World Health Organization. (https://www.who.int/health-topics/social-determinants-of-health#tab=tab_1).

WHO Regional Office for Europe (2023). Population health management in primary health care: a proactive approach to improve health and well-being. Copenhagen: WHO Regional Office for Europe (https://iris.who.int/handle/10665/368805).

Wiemken TL, Kelley RR (2020). Machine learning in epidemiology and health outcomes research. Annu Rev Public Health.41(1):21–36. (https://doi.org/10.1146/annurev-publhealth-040119-094437).

Wiens J et al. (2019). Do no harm: a roadmap for responsible machine learning for health care. Nat Med.25(9):1337–40.

Wilding M (2025). UK police working with controversial tech giant Palantir on real-time surveillance network. Liberty Investigates. 16 June (https://libertyinvestigates.org.uk/articles/uk-police-working-with-controversial-tech-giant-palantir-on-real-time-surveillance-network/).

Williamson SM, Prybutok V (2024). Balancing privacy and progress: a review of privacy challenges, systemic oversight, and patient perceptions in AI-driven healthcare. Appl Sci.14(2):675. (https://doi.org/10.3390/app14020675).

Wilson PW et al. (1998). Prediction of coronary heart disease using risk factor categories. Circulation.97(18):1837–47. (https://doi.org/10.1161/01.cir.97.18.1837).

Winkler-Schwartz A et al. (2019). Machine learning identification of surgical and operative factors associated with surgical expertise in virtual reality simulation. JAMA Netw Open.2(8):e198363. (https://doi.org/10.1001/jamanetworkopen.2019.8363).

Wong SW, Crowe P (2024). Cognitive ergonomics and robotic surgery. J Robot Surg.18(1):110. (https://doi.org/10.1007/s11701-024-01852-7).

Wood M et al. (2024). Paving a new pathway to prevention: leveraging increased returns on our collective investment. London: NHS Confederation (https://www.nhsconfed.org/publications/paving-new-pathway-prevention).

Wornow M et al. (2023). The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit Med.6(1):135. (https://doi.org/10.1038/s41746-023-00879-8).

Wu Y et al. (2024). Application of chatbots to help patients self-manage diabetes: systematic review and meta-analysis. J Med Internet Res.26:e60380. (https://doi.org/10.2196/60380).

Yadav N et al. (2023). Data privacy in healthcare: in the era of artificial intelligence. Indian Dermatol Online J.14(6):788–92. (https://doi.org/10.4103/idoj.idoj_543_23).

Yang D et al. (2025a). Population-wide depression incidence forecasting comparing autoregressive integrated moving average and vector autoregressive integrated moving average to temporal fusion transformers: longitudinal observational study. J Med Internet Res.27:e67156. (https://doi.org/10.2196/67156).

Yang J et al. (2023). An adversarial training framework for mitigating algorithmic biases in clinical machine learning. NPJ Digit Med.6(1):55. (https://doi.org/10.1038/s41746-023-00805-y).

Yang Y et al. (2025b). Demographic bias of expert-level vision-language foundation models in medical imaging. Sci Adv.11(13):eadq0305. (https://doi.org/10.1126/sciadv.adq0305).

Ye Y et al. (2025). Integrating artificial intelligence with mechanistic epidemiological modeling: a scoping review of opportunities and challenges. Nat Commun.16(1):581. (https://doi.org/10.1038/s41467-024-55461-x).

Yu F et al. (2024). Heterogeneity and predictors of the effects of AI assistance on radiologists. Nat Med.30(3):837–49. (https://doi.org/10.1038/s41591-024-02850-w).

Zamzam AH et al. (2023). Integrated failure analysis using machine learning predictive system for smart management of medical equipment maintenance. Engi Appl Artif Intell.125:106715. (https://doi.org/10.1016/j.engappai.2023.106715).

Zewe A (2025). Explained: generative AI's environmental impact. MIT news. 17 Jan (https://news.mit.edu/2025/explained-generative-ai-environmental-impact-0117).

Zhan H (2025). The potential of large language models to achieve artificial general intelligence. Topoi. (https://doi.org/10.1007/s11245-025-10207-2).

Zhang K et al. (2023). Using deep learning to predict survival outcome in non-surgical cervical cancer patients based on pathological images. J Cancer Res Clin Oncol.149(9):6075–83. (https://doi.org/10.1007/s00432-022-04446-8).

Zhang K et al. (2025). Artificial intelligence in drug development. Nat Med.31(1):45–59. (https://doi.org/10.1038/s41591-024-03434-4).

Zhang T et al. (2020). BERTScore: evaluating text generation with BERT. arXiv.1904.09675. (https://doi.org/10.48550/arXiv.1904.09675).

Zhou J et al. (2024a). Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. Nat Commun.15(1):5649. (https://doi.org/10.1038/s41467-024-50043-3).

Zhou L et al. (2024b). Larger and more instructable language models become less reliable. Nature.634(8032):61–8. (https://doi.org/10.1038/s41586-024-07930-y).

Zhou S et al. (2025). Large language models for disease diagnosis: a scoping review. NPJ Artif Intell.1(1):9. (https://doi.org/10.1038/s44387-025-00011-z).

This book is a practical guide for anyone working in or alongside the health sector who needs to engage with the fast-evolving topic of artificial intelligence (AI). It aims to provide a balanced foundation for informed decision-making, helping readers ask the right questions and shape a future where AI serves public health goals. For newcomers, it offers an accessible introduction; for experienced professionals, it aims to broaden perspectives and uncover new opportunities.

Starting with the fundamentals of what AI really is (beyond the hype), it explores diverse approaches, from traditional machine learning to generative models, and how they are transforming (or could transform) clinical care, public health, scientific research and operational efficiency. It reveals both promise and pitfalls: improved outcomes and efficiency on the one hand, and risks such as bias, safety concerns and governance challenges on the other.

Written for readers with varying technical backgrounds, it aims to deliver clear insights into ethical use, trust-building and regulatory frameworks, making it essential reading for health professionals, managers, policy-makers and researchers.

The authors

Paula del Rey Puech, London School of Hygiene & Tropical Medicine and Royal Free London NHS Trust

Jasjot Saund, Royal Free London NHS Trust and AI Centre for Value Based Healthcare

Dimitra Panteli, European Observatory on Health Systems and Policies

Martin McKee, European Observatory on Health Systems and Policies

Health Policy Series No. 63

European Observatory on Health Systems and Policies
a partnership hosted by WHO